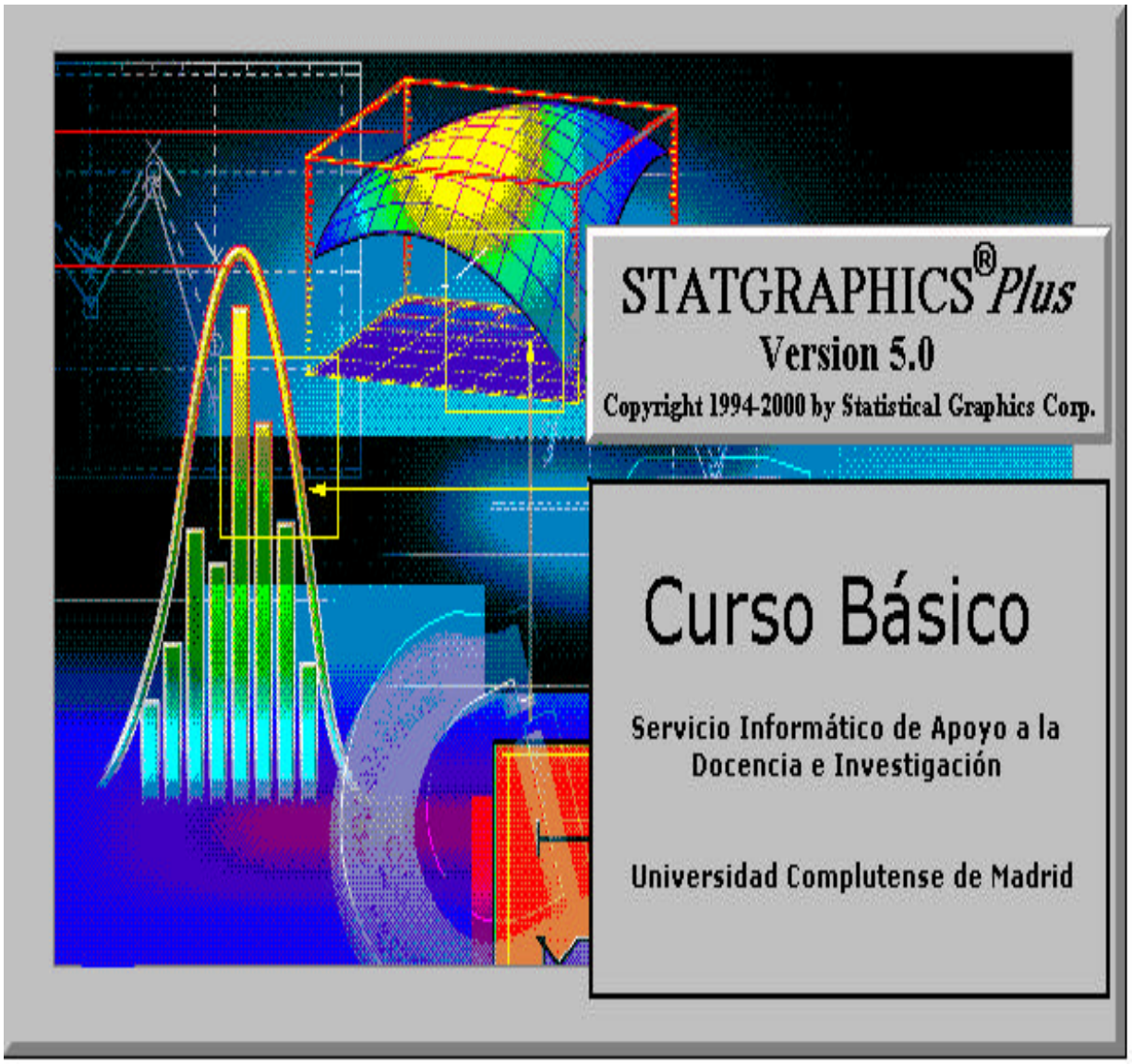

STATGRAPHICS® Plus
Version 5.0
Copyright 1994-2000 by Statistical Graphics Corp.
Curso Básico
Servicio Informático de Apoyo a la Docencia e Investigación
Universidad Complutense de Madrid

ÍNDICE

0. INTRODUCCIÓN

Es cada vez más generalizada la utilización de técnicas estadísticas para el análisis de datos por parte de profesionales e investigadores provenientes de todos los ámbitos. Por otro lado, la complejidad de los cálculos implícitos en muchos modelos matemáticos, así como la creciente necesidad de manipulación de grandes volúmenes de datos, hacen del ordenador una herramienta imprescindible en el análisis estadístico.

Es lógica, por lo tanto, la aparición de programas que permiten la realización sistemática, en un entorno común, de distintos análisis estadísticos a partir de un sistema de comunicación con el ordenador conciso y sencillo para el usuario no especializado.

Entre ellos está STATGRAPHICS, un paquete de *software* para ordenadores personales dirigido por menús que integra una gran variedad de análisis estadísticos y gráficos de alta resolución.

Este curso trata la Versión 5.0 PLUS para WINDOWS (95, 98, 2000, NT 4.0 o superior), que requiere un procesador Pentium y 32Mb de memoria RAM. En la *Figura 1* podemos ver la ventana principal de la aplicación.

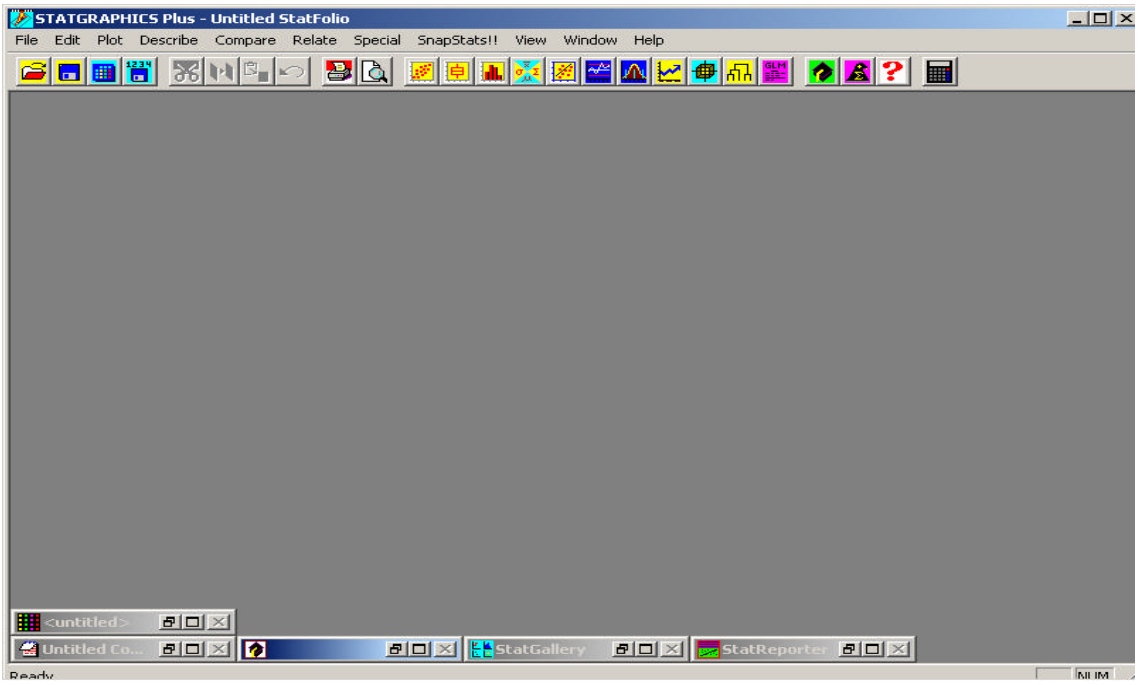


Figura 1

Las diferentes funciones y procedimientos se encuentran accesibles desde las opciones de la Barra de Menú. Durante el curso, trataremos la creación, importación y manipulación de ficheros de datos, representación de gráficos, generación de estadísticas descriptivas, estimación mediante intervalos de confianza y contrastes de hipótesis, y la aplicación de dos de los modelos estadísticos más utilizados: Análisis de la Varianza y Regresión Lineal.

El primer capítulo aborda las generalidades del sistema, los aspectos comunes a las diferentes funciones. En los siguientes, se estudiarán cada una de ellas por separado.

1. GENERALIDADES

Entrada y salida del sistema.

Para entrar en STATGRAPHICS *Plus* debemos invocar el programa SGWIN.EXE, bien ejecutándolo explícitamente o haciendo doble *click* sobre el icono correspondiente. Para salir, seleccionar FILE…EXIT STATGRAPHICS en la Barra de Menú (ver siguiente sección) o, sencillamente, cerrar la ventana principal de la aplicación.

La Ventana Principal de la aplicación.

Al entrar en la aplicación aparecerá la ventana principal que vimos en la *Figura 1*, sobre la que trabajaremos mientras dure la sesión. Distinguimos en ella tres elementos que nos permitirán comunicarnos con Statgraphics para realizar nuestros análisis: la Barra de Menú, la Barra de Herramientas y la Barra de Tareas. Describimos a continuación cada uno de estos tres elementos.

Como parte de la ventana principal de la aplicación, la Barra de Menú siempre estará disponible para seleccionar la función o análisis deseados. Consta de diez palabras clave sobre las que podemos picar con el ratón (*Figura 2)*. Al hacerlo, se nos mostrará un menú emergente con las opciones asociadas. Algunas de ellas (las marcadas con) despliegan a su vez un submenú con nuevas opciones finales.

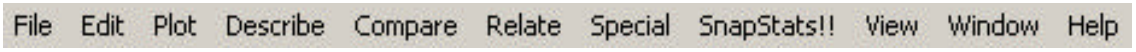


Figura 2

- *File:* Las opciones de este menú permiten realizar operaciones de carácter general como abrir (*Open*), cerrar (*Close*) o grabar (*Save, Save as*) ficheros, imprimir (*Print*) o salir del sistema (*Exit Statgraphics*) entre otros.
- *Edit:* Como en otras aplicaciones en entorno *Windows*, este menú da acceso a diferentes opciones de edición: deshacer la última acción (*Undo*), copiar, cortar y pegar (*Copy, Cut y Paste*) y otras.
- *Plot, Describe, Compare, Relate* y *Special:* Dan acceso a los diferentes análisis estadísticos incorporados en STATGRAPHICS. Veremos algunos de ellos en los capítulos posteriores. Los análisis asociados a *Special* (Control de Calidad, Diseño de Experimentos, Análisis de Series Temporales, Métodos Multivariantes y Regresión Avanzada) quedan fuera del objetivo de este curso.
- *SnapStats!:* Para generar en un único paso varias salidas asociadas a algunos análisis considerados como habituales.
- *View, Window* y *Help*: Proporcionan funciones de formato y ayuda de manera similar a otras aplicaciones en este entorno.

La Barra de Herramientas, que aparece en la *Figura 3*, simplemente asocia iconos con algunas de las opciones más habituales de la barra de menú para proporcionar un acceso más cómodo a las mismas. Señalando cualquiera de los iconos con el ratón aparecerá una breve descripción de la función asociada en el borde inferior de la ventana principal de la aplicación.

Figura 3

La Barra de Tareas (*Figura 4)* contiene iconos asociados a *sub-ventanas* que contendrán elementos diversos como: ficheros de datos, resultados de análisis efectuados sobre ellos, comentarios personales e interpretaciones del sistema sobre dichos resultados, y otros. Todos estos elementos formarán, conjuntamente, lo que conoceremos más adelante por el nombre de *StatFolio*.

Figura 4

El Asesor Estadístico (StatAdvisor)

Es una herramienta incorporada en STATGRAPHICS que produce de manera automática una interpretación corta y fácilmente comprensible de los resultados de los análisis estadísticos realizados.

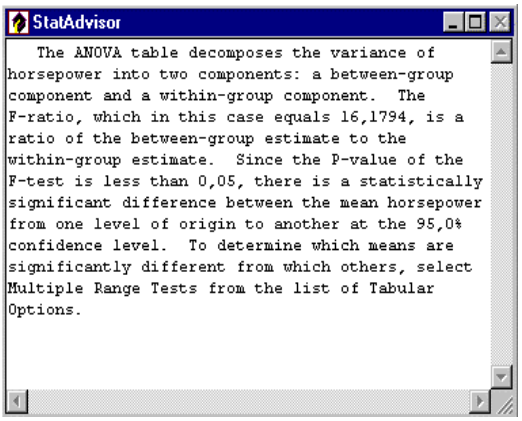


Figura 5: StatAdvisor interpreta una tabla de Análisis de la Varianza y aconseja sobre posteriores análisis

Podemos acceder a esta herramienta desde los iconos correspondientes en la Barra de Herramientas o en la de Tareas. Veremos así la interpretación correspondiente al último análisis realizado. Si queremos guardar la información generada para cada análisis junto con el resultado de los mismos activaremos la opción VIEW…STATADVISOR de la barra de menú.

La Galería de Resultados (StatGallery)

Esta herramienta permite el almacenamiento de los resultados de uno o varios análisis estadísticos, generando así una presentación organizada y personalizada de los mismos. La ventana de *StatGallery* se compone de páginas cada una de las cuales contiene 9 paneles organizados con estructura matricial (3x3). En ellas se pueden almacenar hasta 100 salidas gráficas y un número ilimitado de salidas de texto provenientes de los análisis ejecutados.

Explicaremos ahora brevemente el uso de *StatGallery*. Cuando hallamos generado un fichero de datos y ejecutado un análisis estadístico sobre él (ya veremos cómo realizar estas operaciones en capítulos posteriores), el sistema generará una ventana con los resultados de dicho análisis. Esa ventana estará dividida en paneles que contendrán resultados gráficos y textuales según las opciones que hayamos decidido aplicar.

Pulsando el botón derecho del ratón en uno de estos paneles aparecerá un menú emergente en el que seleccionaremos la opción *Copy to Gallery*. Posteriormente abrimos la ventana de *StatGallery* y pulsamos el mismo botón en el panel en el que queremos cargar la

información. Elegimos esta vez la opción *Paste* y la copia queda realizada. Si elegimos la opción *Paste link* se lleva a cabo una copia dinámica o vínculo, de manera que si realizamos cualquier modificación sobre el panel original en la ventana de análisis correspondiente, dicha modificación se cargará automáticamente en StatGallery sin tener que repetir el proceso de copia.

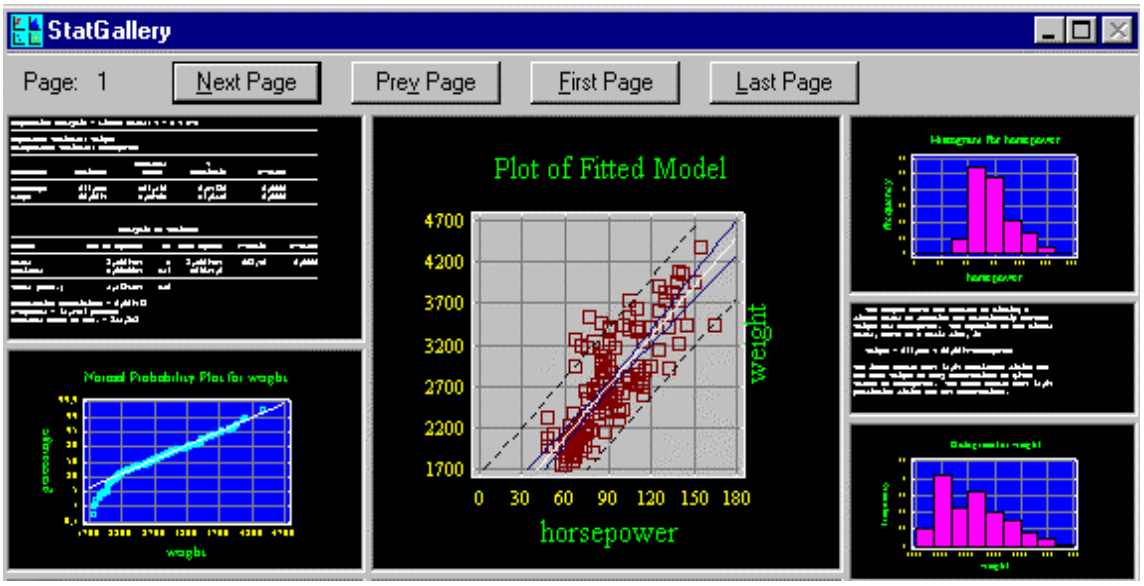


Figura 6

La configuración de paneles dentro de una página de *StatGallery* puede personalizarse sencillamente desplazando con el ratón las barras horizontales y verticales que los delimitan. En la *Figura 6* podemos ver un ejemplo de una página personalizada cargada con distinta información gráfica y textual.

El Editor de Resultados (StatReporter)

A caballo entre un sencillo *notepad* y un completo procesador de textos, esta herramienta nos permitirá generar informes más personalizados que StatGallery, combinando salidas textuales y gráficas generadas por Statgraphics con nuestros propios textos.

Para utilizarlo, sencillamente maximizar la ventana correspondiente en la barra de tareas y escribir como en un procesador de textos. Para incluir análisis generados por el sistema accederemos a la ventana que los contiene y, después de pulsar el botón derecho del ratón, seleccionamos *Copy Analysis to StatReporter*. Para copiar resultados parciales, usaremos *Copy* en la ventana de análisis tras seleccionar la parte deseada, y pegaremos con *Paste* en StatReporter.

El Editor Web (StatPublish)

Nueva herramienta en la versión 5.0 que permite guardar el StatFolio en formato HTML para ser publicado en la Web. Para hacerlo, simplemente seleccionar StatPublish en el menú File. Nos aparecerá un cuadro de diálogo donde podremos indicar qué información guardar, con qué nombre, y opciones de tamaño y formato para los gráficos

El concepto de StatFolio

Aunque ya lo hemos apuntado con anterioridad, describimos ahora con detalle el concepto de *StatFolio*, una de las características más interesantes de STATGRAPHICS.

En la barra de tareas tendremos siempre cinco ventanas activas desde que comencemos la sesión. Tres se corresponden con el *StatAdvisor*, el *StatGallery* y el *StatReporter* que ya hemos comentado en los apartados anteriores. Otra, que en principio aparece con el nombre de *<Untitled>*, contendrá el fichero con los datos a analizar. La quinta (*Untitled Comments*) está ideada para contener comentarios personales del usuario.

Como también hemos comentado anteriormente, generaremos un fichero de datos (o abriremos uno ya generado) y efectuaremos sobre él uno o varios análisis. Cada uno de estos análisis da lugar a una nueva ventana en la barra de tareas que contiene los resultados del mismo.

Llamaremos *StatFolio* al conjunto formado por todos los elementos nombrados anteriormente (StatAdvisor, StatGallery, comentarios personales y resultados de los análisis) que aparecerán en todo momento diferenciados en la barra de tareas.

Podremos almacenar el conjunto bajo un único nombre (**.sgp*) utilizando la opción FILE…SAVE (AS)… SAVE STATFOLIO (AS) de la barra de menú. Para abrir un *StatFolio* ya existente usaremos FILE… OPEN… OPEN STATFOLIO. Ambas funciones tienen asociado un icono en la barra de herramientas. A grabar un *StatFolio* el sistema nos pedirá confirmación para grabar también *StatGallery* (si no está vacío). En caso de contestar afirmativamente, se grabará en un fichero separado con extensión *sgg*. Al abrir un StatFolio el sistema reconocerá y abrirá automáticamente, si existe, el *StatGallery* asociado.

Si después de abrir un *StatFolio* realizamos alguna modificación sobre el fichero de datos, todos los análisis afectados se recalcularán automáticamente. Así podremos corregir errores cómodamente sin repetir todo el proceso. Además, la versión 5.0 incorpora una nueva herramienta que permite crear un vínculo dinámico entre un StatFolio y una fuente de datos externa. Se llama *StatLink* y se accede desde el menú *File*.

También es muy corriente la necesidad de repetir periódicamente una cadena de análisis sobre distintos conjuntos de datos con la misma estructura (p.e. datos mensuales o anuales). Una vez realizado el proceso para el primer conjunto de datos y guardado el conjunto como StatFolio, sólo tendremos que sustituir el fichero en posteriores ocasiones para tener los resultados actualizados.

Vemos que la principal ventaja que nos aporta el *StatFolio* es *la posibilidad de ejecutar sistemáticamente un conjunto de análisis sobre distintos conjuntos de datos sin tener que repetir el proceso ni que programar macros.*

2. GENERACIÓN DE FICHEROS

Conceptos

El conjunto de datos que queramos analizar con STATGRAPHICS ha de ser almacenado en el sistema de una manera lógica y ordenada que permita su reconocimiento y análisis. A estos conjuntos de información almacenados de manera adecuada para STATGRAPHICS los llamaremos *ficheros de datos*.

Llamaremos *variable* a un conjunto de mediciones de la misma característica en determinados individuos de una población. Las variables se agrupan en columnas para formar los ficheros de datos. Un fichero de datos es, por lo tanto, un conjunto de información estructurada en forma matricial que contiene los valores de una o varias características medidas en determinados individuos de una población. Las columnas del fichero (variables) representan los valores de una característica a lo largo de todos los individuos observados, mientras que las filas representan los valores de todas las variables medidas en cada individuo. Es habitual llamar a estas filas *registros* u *observaciones*.

$\mathbf{var_1}$	$\mathbf{var_2}$	...	$\mathbf{var_n}$	
C11	C12	...	C1n	$\mathbf{reg_1}$
C21	C22	...	C2n	$\mathbf{reg_2}$
...	...	...	...	.
...	...	...	...	.
...	...	...	...	.
Ck1	Ck2	...	Ckn	$\mathbf{reg_k}$

Figura 7: Estructura matricial de un fichero de datos con *n* variables medidas en *k* individuos. Cada elemento Cij del fichero representa el valor de la variable *j* en el individuo *i*.

De acuerdo con la naturaleza de las características observadas podemos clasificar las variables según la siguiente tipología:

Variables Categóricas: Son aquellas que reflejan características cuyos valores no admiten una representación numérica con sentido pleno. El número de valores que pueden toman es normalmente pequeño (siempre finito) dividiendo así a la población bajo estudio en clases o *categorías*. Pueden subdividirse a su vez en:

Nominales: Sus valores son meros nombres que no admiten ningún tipo de interpretación numérica. Por ejemplo, el sexo, la raza o la religión. Pueden codificarse como números por comodidad (p.e. 1=Católico, 2=Protestante, 3=Musulmán, 4=Otras) pero la asignación de números a categorías es totalmente arbitraria.

Ordinales: Las categorías que representan admiten una ordenación natural. Por lo tanto admiten una representación numérica donde sólo tendrá sentido interpretar el orden relativo entre los números que representan a distintas categorías. Así podemos recoger la opinión de distintos consumidores sobre una nueva marca en el mercado con una variable que tome los valores 0 (=Nada satisfactorio), 1 (=Poco satisfactorio), 2 (=Satisfactorio) y 3 (=Muy satisfactorio).

Variables Numéricas: Sus valores reflejan cantidades que admiten una representación numérica con sentido pleno. Se subdividen en:

Discretas: La cantidad de valores distintos que pueden tomar es numerable (esto no significa finito, aunque habitualmente lo sea). Los distintos valores son unidades separadas (como categorías, pero con sentido numérico pleno). Como ejemplos, el año de construcción de los edificios o el número de hijos de las familias.

Continuas: Pueden tomar infinitos valores en un intervalo continuo. La altura, el peso o el nivel de colesterol son ejemplos de variables continuas.

STATGRAPHICS sólo puede trabajar con ficheros que estén almacenados en un formato propio (**.sf3*). Existen dos formas de crear estos ficheros: desde dentro de STATGRAPHICS usando su propio editor, o importando ficheros ya creados por otras fuentes. A continuación detallamos ambos métodos.

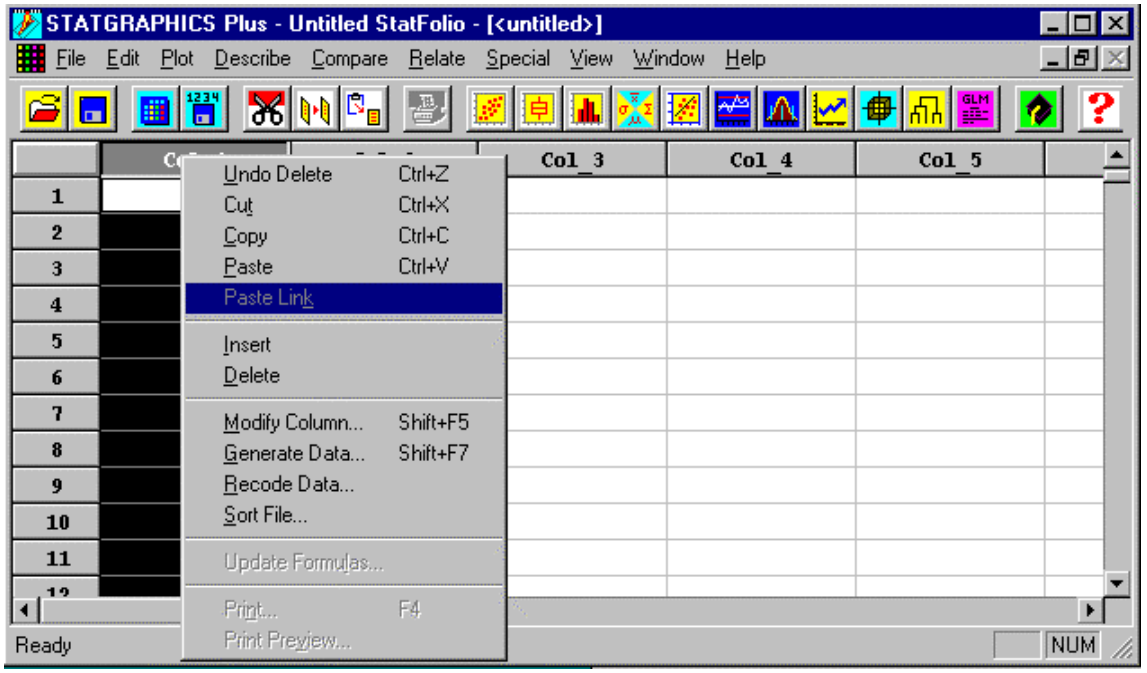


Figura 8

Creación de ficheros con el Editor de Datos

Llamamos *editor de datos* a la herramienta utilizada para generar y editar (visualizar y/o modificar) los ficheros nombrados en el apartado anterior. Para acceder al editor, simplemente maximizaremos la ventana correspondiente al fichero de datos en la barra de tareas que, como ya dijimos, aparecerá al comenzar la sesión bajo el nombre de *<Untitled>*

Como podemos ver en la *Figura 8* el aspecto del editor es muy similar al de las hojas de cálculo más habituales. Las filas (registros) vienen numeradas y las columnas (variables) aparecen con una cabecera sombreada para contener el nombre (por defecto *Col_1, Col_2, …*).

Antes de empezar a grabar los datos, debemos formatear las columnas indicando al sistema cierta información sobre las variables que van a contener. Para ello marcamos una columna picando en su cabecera con el botón izquierdo del ratón. Después picamos con el botón derecho y aparecerá un menú con posibles acciones a aplicar sobre la zona marcada (*Figura 8*). Elegimos *Modify Column* y aparecerá la ventana de la *Figura 9* donde debemos rellenar los siguientes campos:

* *Name:* Nombre que queremos asignar a la variable. Puede tener hasta 32 caracteres. Debe comenzar por una letra (o por uno de estos dos símbolos: _ #). No se distingue entre mayúsculas y minúsculas y se ignoran los espacios en blanco intercalados. Por lo tanto, las variables *E. Civil* y *e.civil* se considerarán equivalentes. La nueva versión 5.0 acepta caracteres

no sajones com la "ñ" o vocales acentuadas. Dos variables del mismo fichero no pueden tener el mismo nombre.

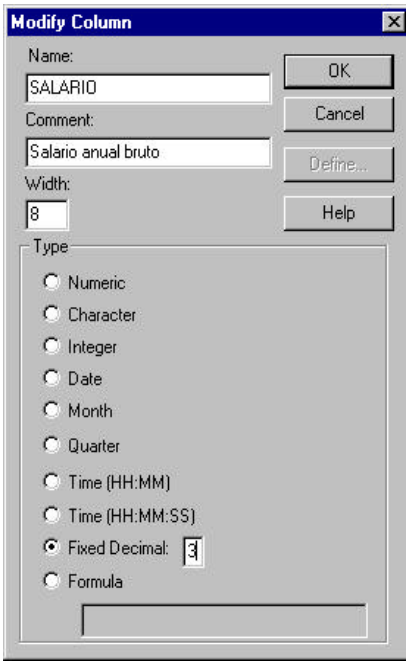


Figura 9

* *Comment:* Escribiremos aquí un comentario personal sobre el contenido de la variable. Puede ser útil para futuras consultas, sobre todo si las hacen usuarios distintos del que creó el fichero.

* *Width*: Anchura de la variable. Debe permitir contener el valor más grande.

* *Type*: Tipo de valores que va a contener la variable.
-*Numeric* para valores numéricos con cualquier tipo de formato (con o sin posiciones decimales).
-*Character* para valores alfanuméricos. Admite cualquier carácter. Es el tipo adecuado para variables nominales
-*Integer* para valores numéricos sin parte decimal
-*Fixed Decimal* para valores numéricos con un número fijo de decimales (que se indica en la casilla adyacente).
-*Date,Month,Quarter,Time* para formatos de fecha y hora
-*Formula* para calcular los valores automáticamente a partir de los valores de otras variables. Hablaremos con más detalle de esta opción más adelante.

Una vez hemos formateado todas las columnas, podemos empezar ya a introducir los datos. Nos posicionaremos en una casilla marcándola con el ratón y escribiremos el valor correspondiente usando el teclado de manera habitual. Para movernos de unas casillas a otras usaremos las teclas de control de movimiento de cursor.

Cuando hayamos terminado, debemos siempre guardar nuestros datos en disco. Para ello, seleccionamos en la barra de menú FILE…SAVE AS…SAVE DATA FILE AS y aparecerá una ventana donde podremos indicar el directorio y nombre elegidos. El fichero se graba y permanece en la ventana del editor, cuyo título cambiará de *<Untitled>* a *nombre_de_fichero.sf3*. Nuestros datos están ya preparados para ser analizados por STATGRAPHICS (*Figura 10)*.

Apertura e importación de ficheros

Si ya tenemos el fichero creado en una sesión anterior y queremos realizar nuevos análisis no es necesario, lógicamente, que volvamos a grabar los datos de nuevo. Sólo tendremos que abrir el fichero y se cargará en la ventana del editor.

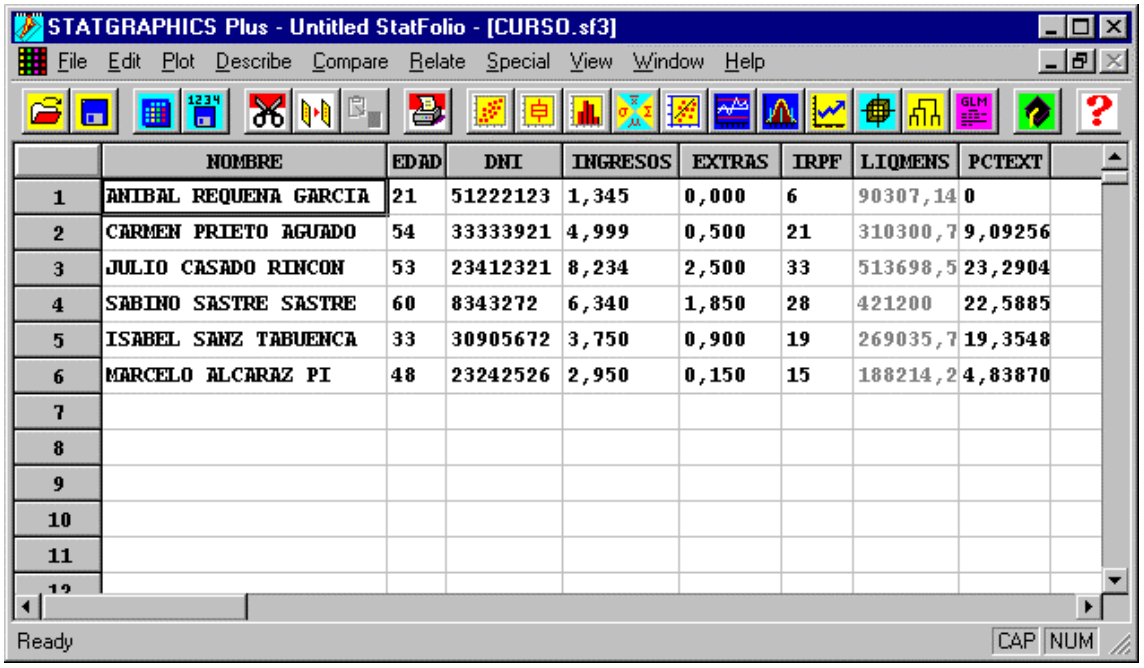
Puede que tengamos nuestros datos en soporte magnético, pero grabados en una fuente externa a STATGRAPHICS, por lo que éste no puede usarlos directamente. Debemos transformarlos al formato propio de STATGRAPHICS, evitando así tener que teclearlos de nuevo (este proceso se llama *importación* de datos).

Para llevar a cabo una de estas acciones seleccionamos FILE…OPEN…OPEN DATA FILE en la barra de menú. Aparece una ventana donde indicaremos el directorio donde se encuentra el fichero que nos interesa. Observamos un campo con el título *Archivos de tipo* que tiene por defecto el valor *SG Plus Files (*.sf3;*.sfx;*.sf).* Esta opción hace referencia a los ficheros que ya están en formato STATGRAPHICS y pueden ser abiertos y analizados directamente por el sistema, que se muestran en la pantalla. Si queremos abrir uno de ellos no tenemos más que seleccionarlo.

En caso de querer importar un fichero, desplegamos el menú *Archivos de tipo* y elegimos el formato correspondiente a la fuente en que fueron grabados los datos. Los ficheros que existan con dicha extensión en el directorio indicado se mostrarán en la pantalla y podremos seleccionar el que nos interesa. Vemos que los formatos que puede importar STATGRAPHICS son: versiones anteriores del propio STATGRAPHICS (*.asf), DIF (*.dif), dBASE (*.dbf), EXCEL (*.xls), LOTUS (*.wk*), EXECUSTAT (*.edf), XML (*.xml) y ASCII. Para acceder a este último tipo de ficheros seleccionaremos *All Files (*.*)*, y el sistema interpretará como tal cualquier fichero con extensión diferente a las anteriormente nombradas.

Al seleccionar el fichero nos aparecerá una nueva ventana para especificar ciertos detalles de la importación que pueden ser diferentes de unos formatos a otros. Después el fichero se cargará y podrá ser analizado. Es importante remarcar que el fichero importado no se graba a disco en formato STATGRAPHICS de forma automática. Por lo tanto, si no lo grabamos como vimos en la sección anterior tendremos que volver a importarlo en futuras sesiones.

Además STATGRAPHICS incluye un controlador ODBC que nos permitirá acceder a distintos tipos de bases de datos (p.e. ACCESS) dependiendo de la configuración ODBC de nuestra máquina. Importaremos datos de esta forma desde *File✍ Open ✍ Query Database(ODBC)*



	NOMBRE	EDAD	DNI	INGRESOS	EXTRAS	IRPF	LIQMENS	PCTEXT
1	ANIBAL REQUENA GARCIA	21	51222123	1,345	0,000	6	90307,14	0
2	CARMEN PRIETO AGUADO	54	33333921	4,999	0,500	21	310300,7	9,09256
3	JULIO CASADO RINCON	53	23412321	8,234	2,500	33	513698,5	23,2904
4	SABINO SASTRE SASTRE	60	8343272	6,340	1,850	28	421200	22,5885
5	ISABEL SANZ TABUENCA	33	30905672	3,750	0,900	19	269035,7	19,3548
6	MARCELO ALCARAZ PI	48	23242526	2,950	0,150	15	188214,2	4,83870
7								
8								
9								
10								
11								

Figura 10

Cierre de ficheros

Cerrar un fichero supone eliminarlo de la memoria del sistema, sacarlo de la ventana de edición y, por lo tanto, perder la disponibilidad de análisis de sus variables. Si hemos creado un fichero por primera vez con el editor, o si hemos realizado alguna modificación en los datos después de abrirlo, no debemos olvidar grabarlo antes de cerrarlo (el sistema pide confirmación a este respecto si no lo hubiéramos hecho).

El sistema cerrará automáticamente el fichero que esté editado al abrir un nuevo fichero de la manera que vimos en el apartado anterior. También podemos hacerlo explícitamente utilizando la opción FILE…CLOSE…CLOSE DATA FILE de la barra de menú.

3. EDICIÓN DE FICHEROS

Una vez que hemos creado un fichero con datos para analizar, es habitual que necesitemos modificar la información que contiene o simplemente visualizarla. Ya hemos apuntado que el proceso mediante el cual podremos llevar a cabo estas funciones se llama *edición*, y se realizará mediante el editor de datos que ya conocemos.

Operaciones básicas de edición

Editar un fichero supone simplemente abrirlo tal y como vimos en el capítulo anterior. Lo más sencillo que podemos hacer es cambiar algún valor concreto o añadir nuevos registros. Para ello simplemente nos posicionamos en la celdilla correspondiente y tecleamos el nuevo valor como cuando creábamos el fichero por primera vez.

Podemos realizar otras operaciones habituales como cortar, copiar y pegar (*Cut*, *Copy* y *Paste*) sobre un bloque de celdas. Para ello marcaremos dicho bloque arrastrando el ratón con el botón izquierdo apretado hasta marcar el rango deseado. Después pulsamos el botón derecho y aparece el menú que vimos en la *Figura 8*, al que nos referiremos a partir de ahora como *Menú de Edición*, y seleccionamos la opción deseada.

Si el rango que queremos seleccionar es una columna o fila completa, sencillamente hay que pulsar el botón izquierdo del ratón sobre la cabecera correspondiente (la celda sombreada que contiene el nombre de la variable en el caso de la columna o el número de observación en el caso de la fila).

Para eliminar un bloque de células o insertarlas en blanco usaremos respectivamente las opciones *Delete* e *Insert* del menú de edición. Hay que tener cuidado pues, si el rango seleccionado no se corresponde con filas o columnas completas, el uso de estas opciones puede romper la integridad de la información contenida en nuestro fichero de datos.

Por último, la opción *Undo* del menú de edición permite deshacer la última acción ejecutada. Recordamos que siempre es posible al cerrar el fichero ignorar todos los cambios que hayamos realizado desde la última vez que lo hayamos grabado (o desde su apertura, si no lo hemos grabado durante la sesión)

Ordenación y recodificación

Si queremos ordenar el fichero completo o parte de él según los valores de una variable concreta seleccionaremos *Sort* en el menú de edición y nos aparecerá una ventana como la de la *Figura 11*.

En ella indicaremos la variable por la queremos ordenar el fichero (*Primary Column)*, una segunda variable según la cual ordenar las observaciones que tengan el mismo valor para la primaria (*Secondary Column*), el tipo de orden deseado (*Order*), y el rango del fichero sobre el que queremos aplicar la ordenación (*Apply to*).

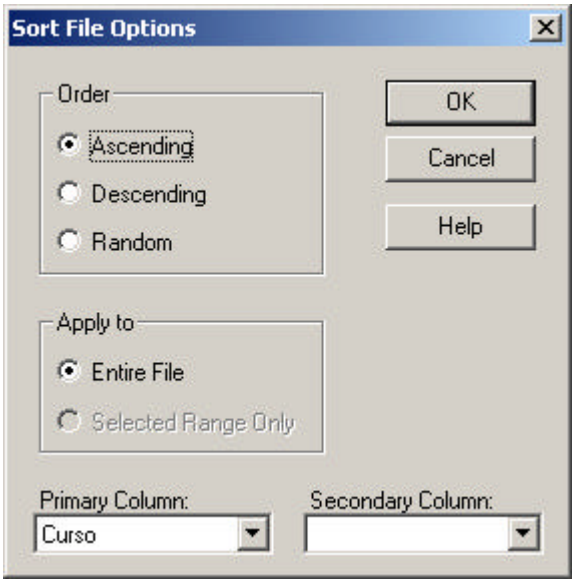


Figura 11

Sobre este último punto es importante remarcar, igual que en el apartado anterior, que ordenar solamente un rango previamente seleccionado del fichero (*Selected Range Only*) puede alterar la integridad de nuestro fichero de datos de manera irreversible.

Si queremos cambiar los valores de una variable de manera que se agrupen aquellos que estén dentro de un mismo rango (intervalo de valores, ya sean numéricos o alfanuméricos), marcamos la variable en el editor y seleccionamos *Recode Data* en el menú de edición.

Nos aparecerá una ventana como la de la *Figura 12* en la que indicaremos los extremos de los intervalos (*Lower Limit* y *Upper Limit*) y el nuevo valor (*New Value*) que queremos asignar a la variable según el intervalo que contenga el valor original (para variables nominales se utilizará el orden alfabético).

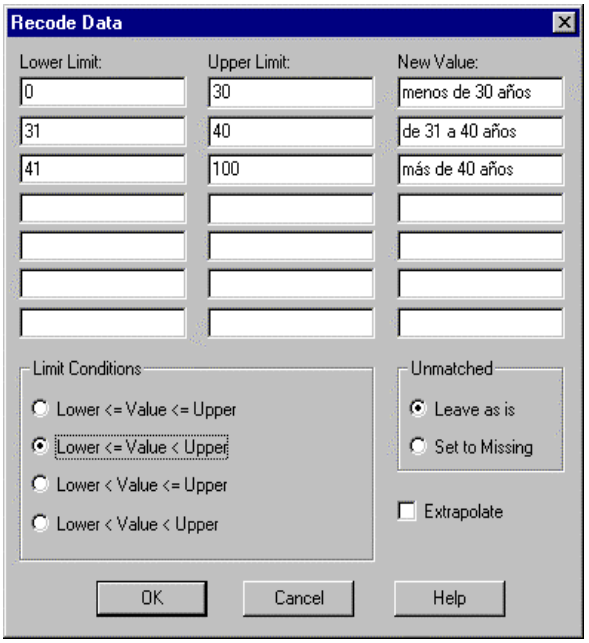


Figura 12

Además debemos indicar el tratamiento deseado para los extremos de los intervalos (*Limit conditions*) y para aquellos valores no contenidos en ninguno de ellos (*Unmatched*). En este último caso podemos conservar los datos originales (*Leave as is*) o bien considerarlos como datos faltantes de manera que las observaciones que los contienen sean eliminadas de cualquier análisis en el que intervenga dicha variable (*Set to missing*).

Esta opción es especialmente útil para la categorización de variables numéricas. En el ejemplo de la *Figura 12* vemos como agrupamos los individuos de nuestro fichero en tres categorías de la variable edad.

Los valores recodificados se cargan en la misma variable, por lo que los valores originales se pierden. Si queremos conservar la variable original junto con la recodificada, duplicaremos la primera (combinando las opciones *Insert*, *Modify column*, *Copy* y *Paste*) y recodificaremos la variable duplicada. Además, el tipo de la variable ha de ser consistente con los valores recodificados. Así, si categorizamos una variable numérica a valores nominales debemos primero transformar el tipo de la variable a carácter (con *Modify column*), o bien duplicarla como explicamos anteriormente.

Generación de variables calculadas

Es muy habitual que queramos añadir una nueva variable al fichero cuyos valores en cada observación puedan ser calculados a partir de los valores de otras variables ya existentes. Sería muy tedioso que tuviéramos que calcular los nuevos datos por nuestra cuenta y grabarlos manualmente. Vemos a continuación cómo el editor de STATGRAPHICS puede hacer el trabajo por nosotros.

En nuestro fichero CURSO de la *Figura 10* tenemos, además de NOMBRE, EDAD y DNI, las variables INGRESOS (que representa los ingresos regulares anuales brutos del individuo en millones de pesetas), EXTRAS (que representa los ingresos no regulares, también en millones de pesetas) e IRPF (que representa el porcentaje de descuento que le corresponde). En realidad, estas son las seis únicas variables que tuvimos que grabar originalmente.

Tenemos además otras dos variables: LIQMENS (que representa el salario mensual líquido del individuo en pesetas) y PCTEXT (que representa el porcentaje que sus ingresos no regulares suponen sobre el total de ingresos).

Para generar la primera procedemos de manera habitual usando la opción *Modify Column* del menú de edición. Tras indicar el nombre y la anchura seleccionamos la opción *Formula* para el campo *Type* y pulsamos con el ratón el botón *Define* (ver *Figura 9*). Aparece entonces la ventana de generación de datos de la *Figura 13*. En el campo *Expression:* escribiremos la fórmula adecuada para generar los nuevos datos a partir de las variables ya existentes. Lo haremos tecleando la expresión directamente, o bien seleccionando con el ratón en los paneles inferiores de la pantalla, donde aparecen todas las variables del fichero y todos los operadores y funciones matemáticas incorporadas en STATGRAPHICS. En la *Figura 13* podemos ver la expresión utilizada para generar la variable LIQMENS.

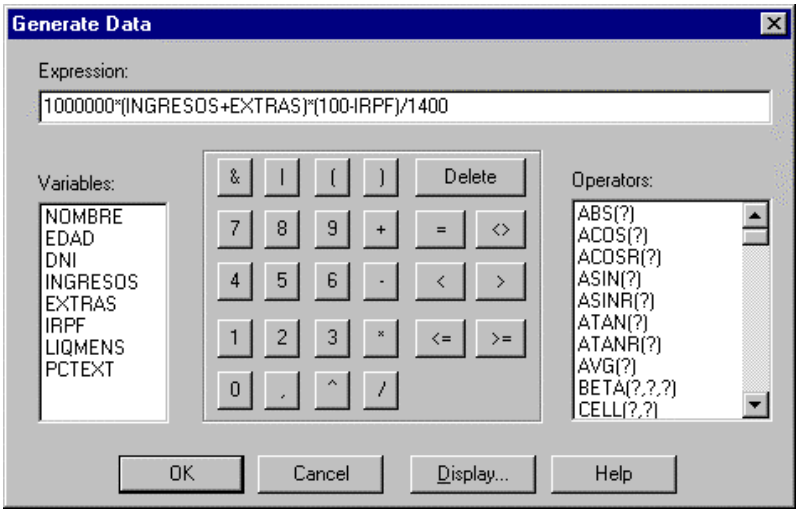


Figura 13

Para generar la segunda, PCTEXT, el proceso seguido es similar. Sin embargo, al indicar el tipo en la pantalla *Modify Column* seleccionamos *Numeric* en vez de *Formula*. Entonces marcamos la variable completa, todavía sin valores, y pulsamos el botón derecho del ratón para que aparezca el menú de edición. Aquí seleccionamos la opción *Generate Data* y aparece de nuevo la pantalla de generación de datos que ya conocemos. En este caso la expresión sería 100*EXTRAS/(INGRESOS+EXTRAS).

Naturalmente nos preguntamos cuál es la diferencia entre ambas formas de generación. En el primer caso (LIQMENS) la variable creada queda vinculada a las originales de manera que la única forma de modificar un valor en ella es modificando alguna de las que intervienen en la expresión de la fórmula que la generó. Para actualizar los valores de la variable calculada después de modificar las originales es necesario aplicar la opción *Update Formulas* del menú de edición. Los valores de las variables generadas de esta manera aparecen en el editor en un tono más claro para indicar que no pueden ser modificados manualmente.

En el segundo caso (PCTEXT) no se establece ningún vínculo posterior al del momento de la generación de la nueva variable. Los valores resultantes pueden ser modificados manualmente y no se pueden actualizar de manera automática a partir de cambios realizados en las originales (salvo, por supuesto, repitiendo el proceso de generación).

Ejecución de procedimientos

Aunque fuera del contexto general del capítulo, abordamos a continuación el tema de la ejecución de los distintos procedimientos estadísticos que incorpora STATGRAPHICS. El desarrollo, interpretación y particularidades de cada uno de ellos se verá en capítulos posteriores. Nos ocupamos ahora de la operativa general, de aquellos aspectos comunes cuyo conocimiento es necesario para la ejecución de todos ellos.

En primer lugar, después de tener abierto nuestro fichero de datos, seleccionaremos el procedimiento o análisis deseado entre los menúes desplegables correspondientes a las opciones *Plot, Describe, Compare, Relate* y *Special* de la barra de menú. Algunos de ellos tendrán asociado un icono en la barra de herramientas que facilita su acceso. Nos aparecerá entonces una pantalla cuyo título será el del análisis elegido y que llamaremos *pantalla de entrada de datos* (*Input dialog*). En la *Figura 14* podemos ver la correspondiente al procedimiento *One Variable Analysis*.

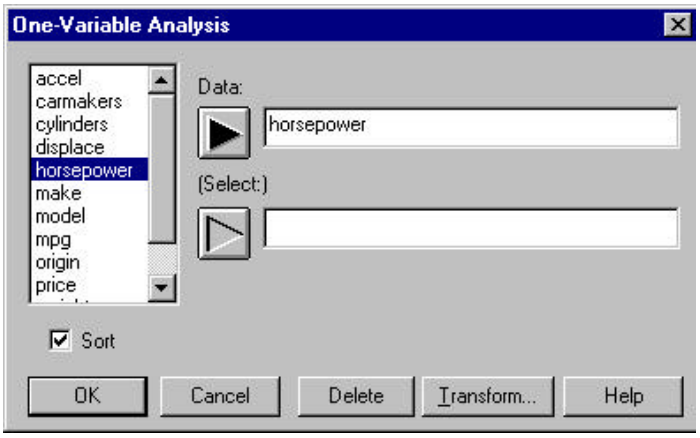


Figura 14

A la izquierda de la ventana tendremos siempre disponible una lista con las variables contenidas en nuestro fichero de datos. A la derecha tendremos distintos campos donde cargar las variables seleccionadas según el análisis elegido. En nuestro ejemplo sólo tenemos un campo obligatorio (*Data:*) donde indicaremos a qué variable queremos aplicar el análisis. Siempre aparecerá el campo opcional *Select:* donde podremos escribir una condición lógica que restrinja nuestro análisis a las observaciones que la cumplan. También podemos utilizar el botón *Transform* para trabajar con transformaciones de las variables originales sin tener que generarlas explícitamente en el fichero de datos como vimos en el apartado anterior. En este caso la variable calculada sólo existirá temporalmente mientras se ejecuta el procedimiento.

Ejecutamos el procedimiento con el botón *OK* y nos aparece una *ventana de análisis* con el título del análisis elegido. En ella aparece una nueva barra de herramientas en la que los tres primeros iconos son de especial relevancia.

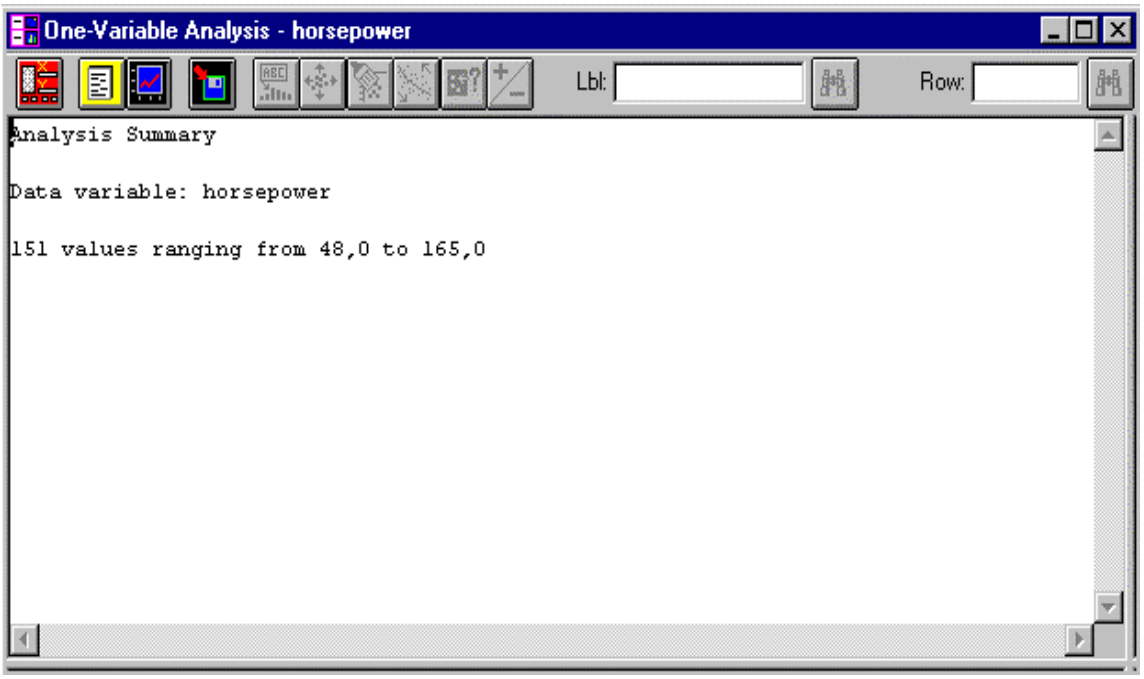


Figura 15

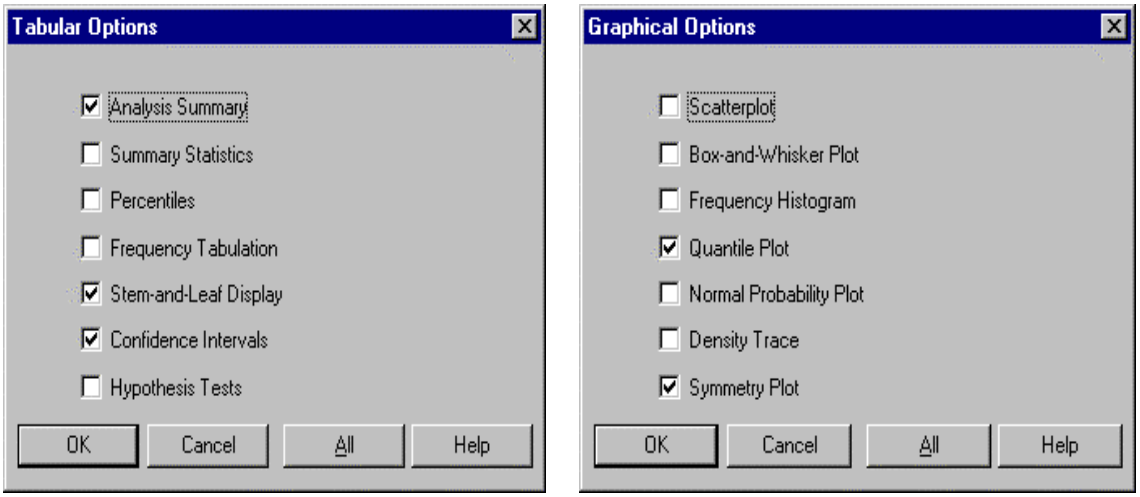


Figura 16

El primero de ellos (*Input Dialog*) vuelve a mostrar en cualquier momento la pantalla de entrada de datos que ya conocemos por si queremos realizar alguna modificación.

Los dos siguientes muestran pantallas con opciones que representan los distintos análisis finales asociados al procedimiento elegido. Los análisis cuyos resultados se muestran en formato textual se agrupan en el icono cuyo rótulo es *Tabular Options*, mientras que aquellos cuya salida se expresa en modo gráfico lo hacen en el correspondiente a *Graphics Options*. En la *Figura 16* vemos todos los análisis finales asocia dos al procedimiento *One Variable Analysis*.

Inicialmente siempre se ejecutará por defecto el *Analisis Summary* que siempre aparecerá como primera opción en *Tabular Options*, cuya salida contiene simplemente información muy general, como cuáles son las variables seleccionadas y el número de observaciones utilizadas (*Figura 15*). El resto de análisis que deseemos debemos seleccionarlos explícitamente.

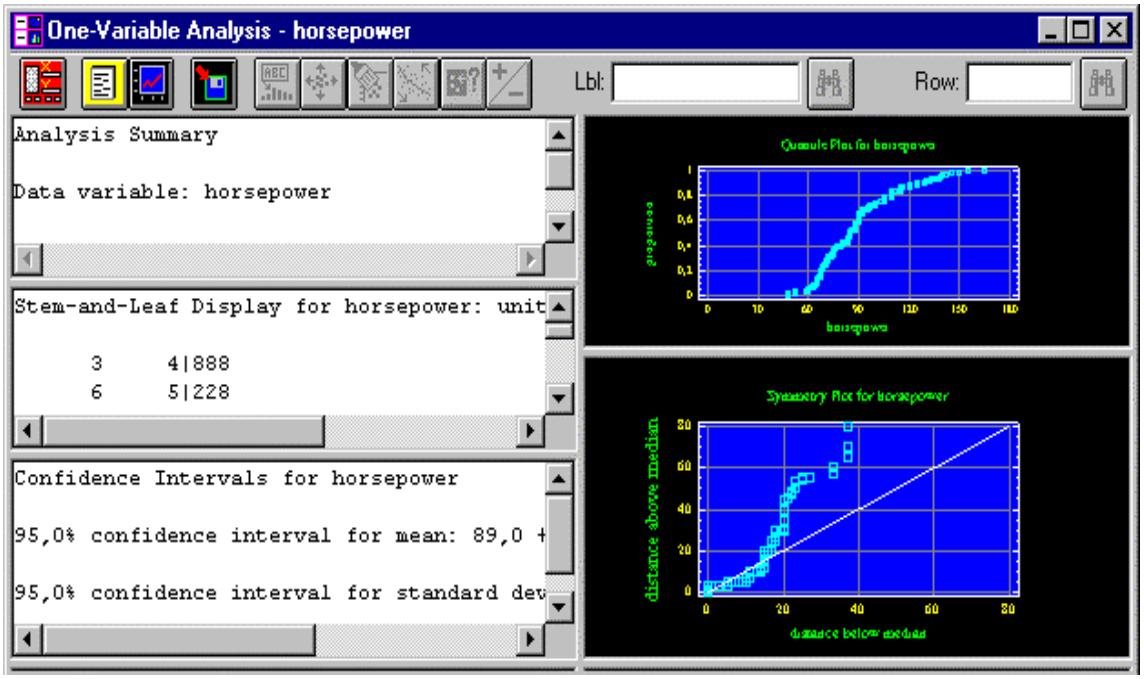


Figura 17

Una vez hecho esto la pantalla del análisis quedará dividida en tantos paneles como análisis finales hayamos seleccionado. Cada panel contendrá los resultados asociados a uno de ellos. En la *Figura 17* vemos la ventana resultante de aplicar todos los análisis indicados en la *Figura 16*.

Como vemos, si seleccionamos varios análisis los paneles resultantes no pueden mostrar los resultados de cada uno de ellos en su totalidad. Para maximizar un panel dentro de la ventana de análisis haremos doble *click* sobre él con el botón izquierdo del ratón. Repitiendo la operación volvemos a la situación original.

Figura 18

Si picamos con el botón derecho del ratón sobre uno de los paneles aparecerá el menú de la *Figura 18* que ofrece distintas opciones.

Entre ellas aparecen opciones de edición (*Undo, Cut, Copy* y *Paste* cuyo uso es igual que en el editor), de impresión (*Print* y *Print Preview*) y de presentación de resultados (*Copy to Gallery*, que ya vimos en el capítulo anterior).

Merecen especial atención las dos opciones que no hemos comentado y que aparecen las primeras en el menú: *Pane Options* y *Análisis Options*.

La primera (*Pane Options*) nos permitirá cambiar algunos parámetros del modelo que, en un principio, están fijados por defecto. Por ejemplo, el número de categorías a generar para la creación de un histograma o el nivel de confianza para la construcción de un intervalo.

La segunda (*Analysis Options*) nos permitirá elegir entre varias características globales que decidirán el modelo final a aplicar. Por ejemplo, el método de selección de variables para un análisis de regresión múltiple. Estas dos opciones no aparecerán activas en todos los paneles y, cuando aparezcan, las distintas alternativas que ofrezcan serán totalmente dependientes del análisis que estemos realizando.

Una vez finalizado el análisis podemos minimizar la ventana correspondiente, empezar otro y así sucesivamente (siempre sobre el mismo fichero de datos). Tendremos al final en la barra de tareas de la ventana general de la aplicación tantos iconos asociados a pantallas de análisis como análisis diferentes hayamos realizado. Podremos entonces imprimir los resultados individualmente o presentarlos de manera conjunta en *StatGallery*. Podemos también, como dijimos en el capítulo anterior, guardar el conjunto bajo un único nombre grabándolo como un *StatFolio*.

4. ESTADÍSTICA DESCRIPTIVA

Conceptos

Las medidas recogidas y grabadas en un fichero de datos constituyen la información básica disponible para el investigador. Sin embargo, la visión conjunta de una gran cantidad de datos no nos permite extraer las características fundamentales del conjunto en sí.

La Estadística Descriptiva trata de mostrar de una manera concisa y resumida los aspectos fundamentales de un conjunto de datos. Esto supone el cálculo de medidas centrales, la cuantificación de la dispersión general de los datos alrededor de las mismas, la presentación resumida de los datos en forma de tablas y gráficos, la detección de datos atípicos, agrupaciones, tendencias, etc.

STATGRAPHICS proporciona un amplio número de métodos descriptivos agrupados bajo la opción DESCRIBE de la barra de menú. El menú asociado, así como los submenúes que nos interesan para el curso, se muestra en la *Figura 19*.

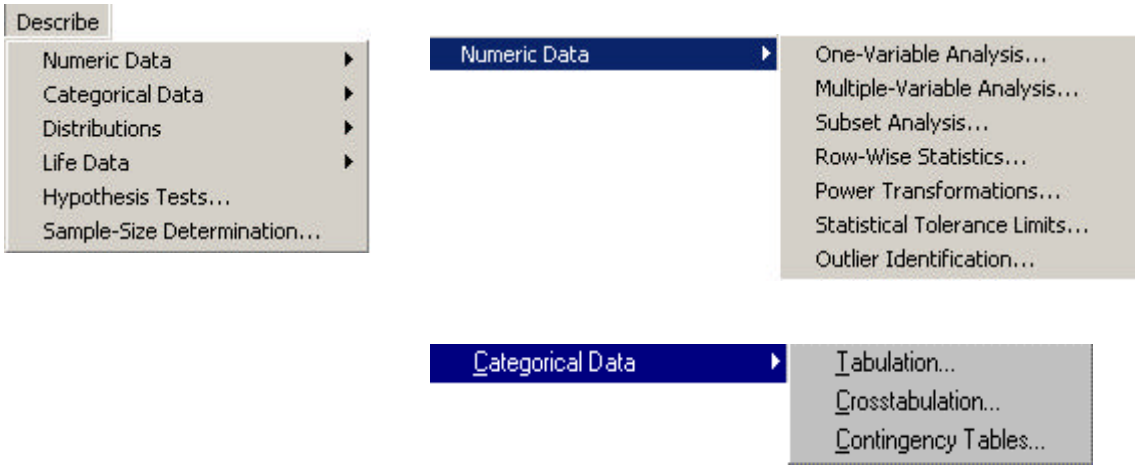


Figura 19

A lo largo del curso trabajaremos habitualmente con el fichero CARDATA, que es uno de los ficheros de ejemplo que vienen incorporados en STATGRAPHICS. Lo encontraremos en *unidad_de_disco*:\SGWIN\DATA y contiene información sobre 155 modelos de automóviles. Sus variables son:

MPG: Consumo en millas/galón
CYLINDERS: Número de cilindros
DISPLACE: Cilindrada en pulgadas cúbicas
HORSEPOWER: Potencia en caballos
ACCEL: Aceleración en segundos de 0 a 60 millas
YEAR: Año del modelo
WEIGHT: Peso en libras
ORIGIN: 1=U.S.A., 2=Europa, 3=Japón
MAKE: Marca
MODEL: Modelo
PRICE: Precio en dólares
CARMAKERS: Etiquetas para la variable ORIGIN

Por lo tanto abrimos este fichero tal y como vimos en el capítulo anterior. Estamos ya en disposición de ejecutar nuestros primeros análisis.

Resumen estadístico

El análisis SUMMARY STATISTICS produce hasta 19 estadísticos asociados a una variable de datos numéricos: media aritmética, varianza, desviación típica, error estándar de la media, mediana, moda, media geométrica, mínimo, máximo, rango, cuartil superior, cuartil inferior, rango intercuartílico, coeficientes (y coef. estandarizados) de simetría y curtosis, coeficiente de variación y suma. Para acceder e él seleccionamos el procedimiento DESCRIBE… NUMERIC DATA… ONE VARIABLE ANALYSIS.

En la pantalla de entrada de datos que ya vimos en el capítulo anterior (*Figura 14)* sólo tenemos que indicar qué variable queremos analizar (*Data:*), por ejemplo HORSEPOWER. Al seleccionar SUMMARY STATISTICS en *Tabular options* se nos muestran los resultados en un panel como el de la *Figura 20*. Por defecto se calculan los estadísticos de uso más común. Podemos añadir o excluir estadísticos del análisis en *Pane Options*. En nuestro ejemplo hemos seleccionado todos menos la media geométrica y los coeficientes de curtosis y simetría.

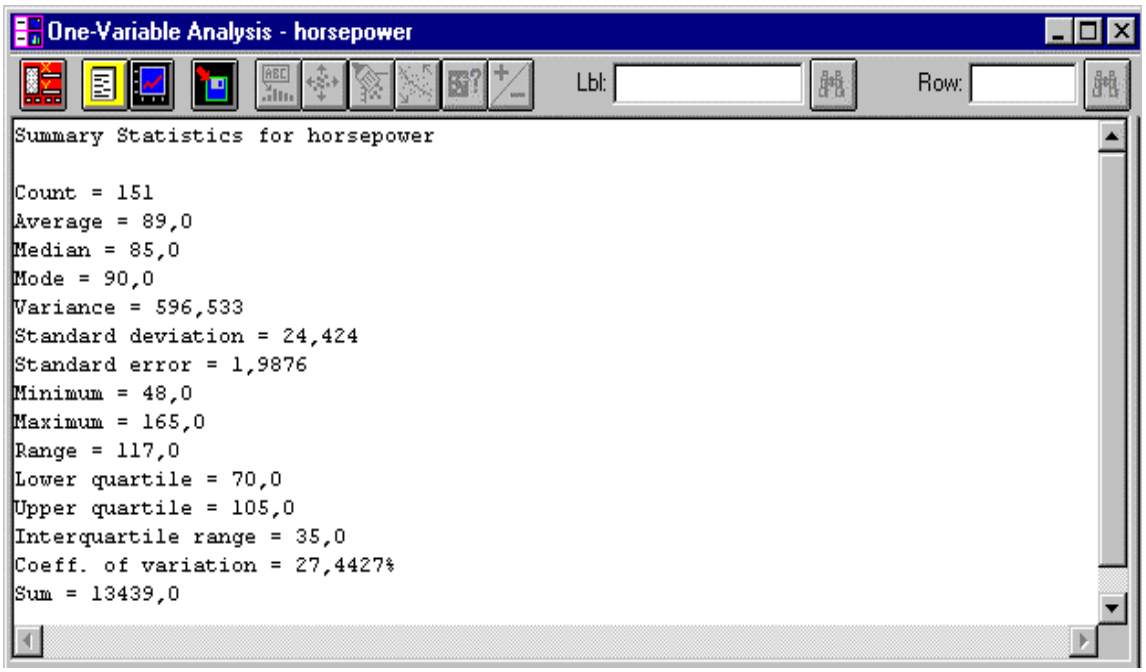


Figura 20

Podemos realizar este análisis de forma simultánea para varias variables si entramos por DESCRIBE… NUMERIC DATA… MULTIPLE VARIABLE ANALYSIS.

Tablas de frecuencias

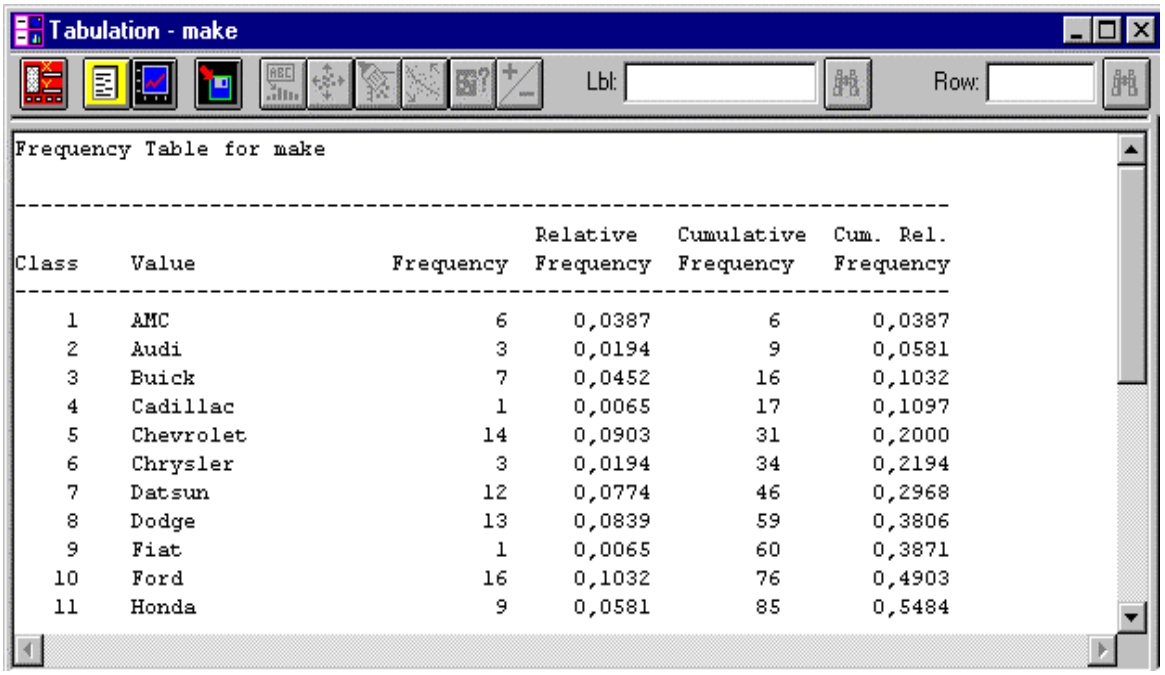
El análisis FREQUENCY TABULATION nos va a permitir resumir la distribución de los datos contenidos en una variable mediante tablas de frecuencias. Para variables discretas y categóricas, estas tablas nos dirán cuántas observaciones tienen cada posible valor de esa variable. Para las continuas se crearán intervalos que constituyan una partición del rango total de la variable, y las tablas nos dirán cuántas observaciones tienen su valor dentro de cada uno de los posibles intervalos.

El número de observaciones en cada valor (o intervalo) se llama *frecuencia absoluta*. El porcentaje que representa sobre el total de las observaciones se llama *frecuencia relativa*. Las *frecuencias acumuladas* representan el numero de observaciones (o porcentaje que representan,

según se trate de frecuencias absolutas o relativas) que toman un determinado valor y todos los menores que él.

En el caso de variables numéricas continuas accederemos por el procedimiento DESCRIBE… NUMERIC DATA… ONE VARIABLE ANALISIS. En *Tabular Options* seleccionamos FREQUENCY TABULATION y, por último, podremos modificar el número de intervalos creados en *Pane Options*.

Para variables categóricas (o numéricas discretas si queremos respetar los valores originales sin agruparlos en intervalos) entramos por el procedimiento DESCRIBE… CATEGORICAL DATA… TABULATION y, de igual manera seleccionamos FREQUENCY TABLE en *Tabular Options* después de indicar la variable a analizar en la ventana de entrada de datos. En la *Figura 21* vemos el resultado de tabular la variable MAKE.



Frequency Table for make

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	AMC	6	0,0387	6	0,0387
2	Audi	3	0,0194	9	0,0581
3	Buick	7	0,0452	16	0,1032
4	Cadillac	1	0,0065	17	0,1097
5	Chevrolet	14	0,0903	31	0,2000
6	Chrysler	3	0,0194	34	0,2194
7	Datsun	12	0,0774	46	0,2968
8	Dodge	13	0,0839	59	0,3806
9	Fiat	1	0,0065	60	0,3871
10	Ford	16	0,1032	76	0,4903
11	Honda	9	0,0581	85	0,5484

Figura 21

Si después de CATEGORICAL DATA seleccionamos CROSSTABULATION podremos obtener la tabulación cruzada de dos variables en una tabla con estructura matricial. La pantalla de entrada de datos nos pedirá qué variables queremos tabular (y su posición en filas o columnas). La tabla resultante tendrá tantas filas como categorías diferentes tenga la variable indicada en *Row Variable:*, y tantas columnas como categorías tenga la indicada en *Column Variable:*.

Cada celda en la tabla contendrá dos números: el número total de observaciones cuyos valores coinciden con las categorías que corresponden a la fila y columna a las que dicha celda pertenece, y el porcentaje que dichas observaciones representan sobre el total. Podemos hacer que dicho porcentaje se refiera al total fila o total columna sin más que seleccionar la opción correspondiente en *Pane Options*. Los totales de fila y columna (que reciben el nombre de *marginales*), con sus correspondientes porcentajes sobre el total, también aparecen en la tabla que nos proporciona así la tabulación individual de cada una de las variables sin que tengamos que usar el procedimiento TABULATION.

Histogramas de frecuencias

Con FREQUENCY HISTOGRAM podremos generar histogramas de frecuencias para variables numéricas. Éstos son representaciones gráficas de las tablas estudiadas en el apartado anterior, en los cuales a cada grupo o intervalo se le asigna una barra cuya altura representa su frecuencia (absoluta, relativa o acumulada).

Encontraremos el análisis en *Graphics Options* del procedimiento DESCRIBE… NUMERIC DATA… ONE VARIABLE ANALYSIS. En *Pane Options* podemos seleccionar el número de intervalos a crear, qué tipo de frecuencia queremos representar y otros parámetros. En la *Figura 22* podemos ver el resultado de analizar la distribución de una variable de tipo continuo, HORSEPOWER, cuyo rango dividimos en siete subintervalos.

Este histograma es la primera salida de tipo gráfico que vemos. Podemos guardarlo en un fichero externo con formato *Windows Metafile (*.wmf)* para incorporarlo posteriormente a otro documento. Para ello usamos la opción FILE… SAVE GRAPH de la barra de menú.

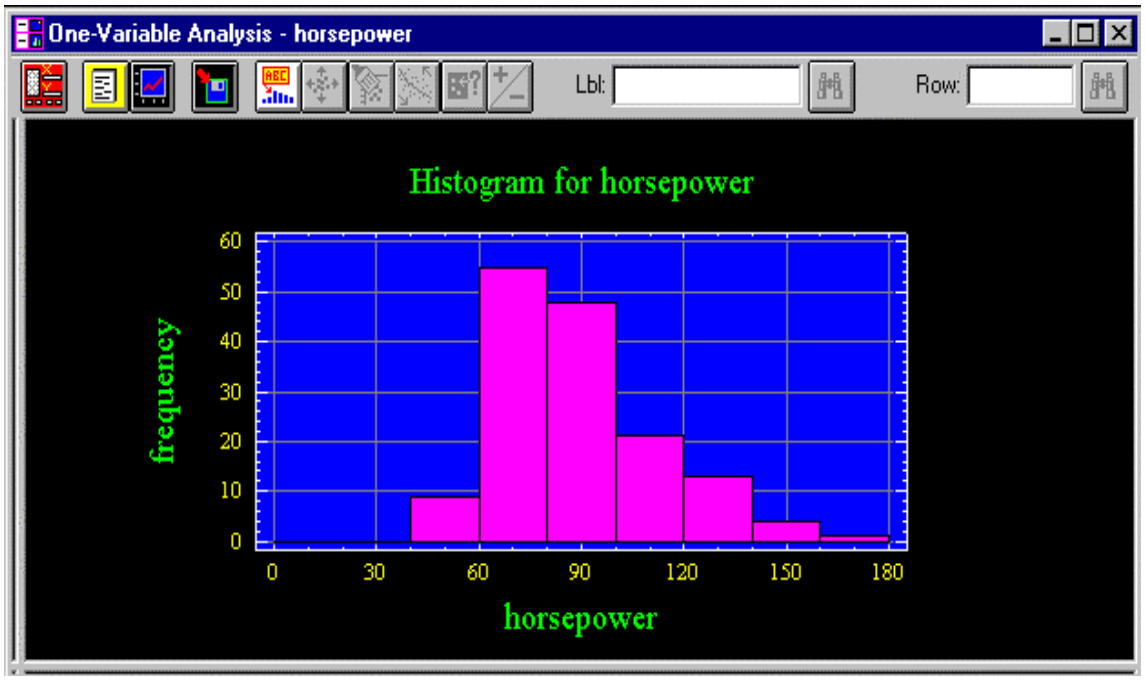


Figura 22

Vemos que el histograma nos da información acerca de la forma de la distribución de la variable analizada. La mayoría de los modelos de coches en nuestro fichero tienen entre 60 y 90 caballos. El resto tiende a tener más, mostrando una distribución no simétrica.

Percentiles

Los *percentiles* son valores que también nos aportan información sobre la distribución de la variable analizada. El *percentil de orden k* de una variable es un valor que es mayor que el k% de los valores que toma la variable, y menor o igual que el (100-k)% restante. Son especialmente utilizados los percentiles de orden 50 (llamado *mediana*), 25 (o *cuartil inferior*), y 75 (o *cuartil superior*).

Hemos visto que estos percentiles especiales son calculados por algún procedimiento de los ya comentados anteriormente. La opción PERCENTILES del *Tabular Options* asociado al

procedimiento DESCRIBE… NUMERIC DATA… ONE VARIABLE ANALYSIS que ya conocemos nos permite calcular los percentiles de cualquier orden de porcentaje.

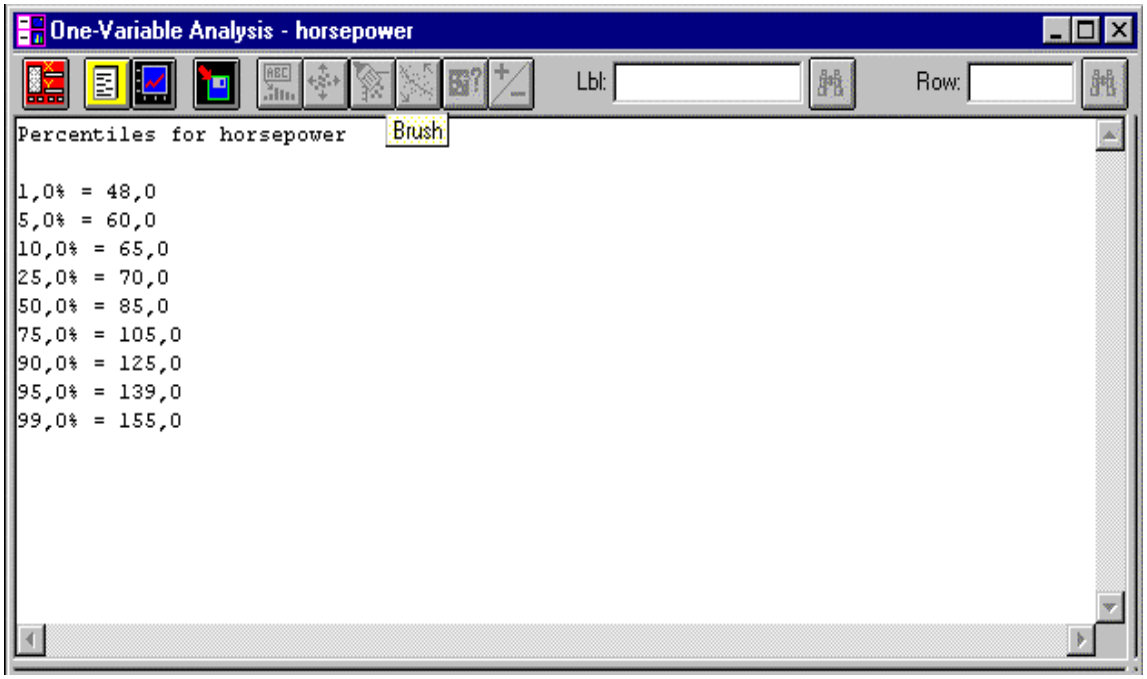


Figura 23

En la *Figura 23* vemos los percentiles calculados por defecto para la variable HORSEPOWER. Si queremos otros diferentes, podemos cambiar los porcentajes correspondientes en *Pane Options*.

Análisis por grupos

El procedimiento DESCRIBE… NUMERIC DATA… SUBSET ANALYSIS, permite realizar distintos análisis descriptivos simultáneamente para distintos subconjuntos de individuos. Los análisis que realiza son el cálculo de los mismos estadísticos que SUMMARY STATISTICS, tablas de medias y una serie de gráficos con medias, intervalos de confianza y errores estándar por grupo. Esto nos permitirá observar las diferencias entre las diferentes medidas centrales y de dispersión de los grupos (OJO!!!, sin que nos permita *inferir* más allá del conjunto de individuos en el que se han recogido los datos).

Los grupos se definirán a partir de los valores de una variable secundaria (*Codes*, en la pantalla de entrada de datos) que pueden etiquetarse (campo *Labels)* .

5. REPRESENTACIÓN GRÁFICA DE DATOS

La representación gráfica de nuestros datos puede facilitarnos información rápida y clara sobre aspectos fundamentales subyacentes en los mismos. STATGRAPHICS incorpora una amplia gama de procedimientos gráficos, agrupados bajo la opción PLOT de la barra de menú. Podemos ver el menú asociado a esta opción en la *Figura 24*. Además, como ya sabemos, encontraremos la posibilidad de realizar gráficos específicos para otros procedimientos en el menú de *Graphics Options* asociado a los mismos.

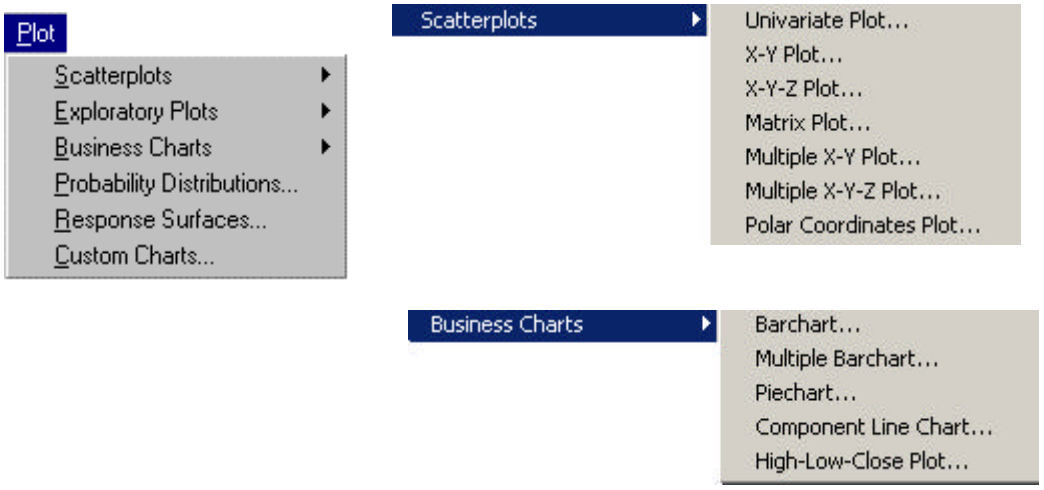


Figura 24

Gráficos de puntos bi/tridimensionales

Con el procedimiento PLOT… SCATTERPLOTS… X-Y PLOT representaremos como puntos en un plano las observaciones de nuestro fichero de datos. Las coordenadas de los puntos serán los valores que dichas observaciones toman para dos de las variables en el fichero.

Podremos así observar si ambas variables presentan un patrón de variación conjunta en las observaciones (esto es, si los individuos que tiene los mayores/menores valores para una de las variables son los mismos que los que los tienen para la otra). También observaremos si existen agrupamientos, tendencias u observaciones muy atípicas con respecto al resto.

El procedimiento incorpora además utilidades como el trazado de líneas que unan los puntos (especialmente adecuado para representación de datos temporales), asignación de códigos a los puntos según los valores de una tercera variable, y otras. Estas opciones estarán accesibles desde la ventana de *Pane Options*.

Vamos a realizar un gráfico de puntos con las observaciones de nuestro fichero CARDATA. Cada punto en el gráfico resultante representará, por lo tanto, un registro de nuestro fichero. Utilizaremos como coordenadas para representar cada automóvil su potencia (HORSEPOWER) en las abscisas y su precio (PRICE) en las ordenadas. Además queremos identificar en el gráfico la procedencia de cada automóvil, esto es, que cada punto se represente con un símbolo que identifique a la variable ORIGIN. La *Figura 25* muestra la pantalla de entrada de datos del procedimiento. Vemos en ella algo nuevo: la manera de restringir el análisis a un subconjunto de observaciones usando el campo SELECT (queremos representar únicamente los modelos que tengan 4 cilindros). La *Figura 26* muestra cómo representar en el mismo gráfico la variable ORIGIN usando *Pane Options*. En la *Figura 27* podemos ver el gráfico resultante.

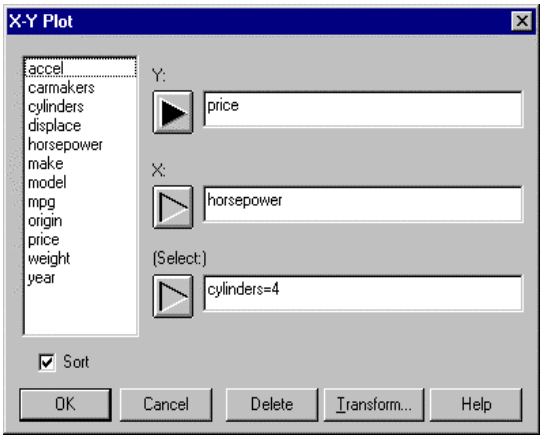


Figura 25

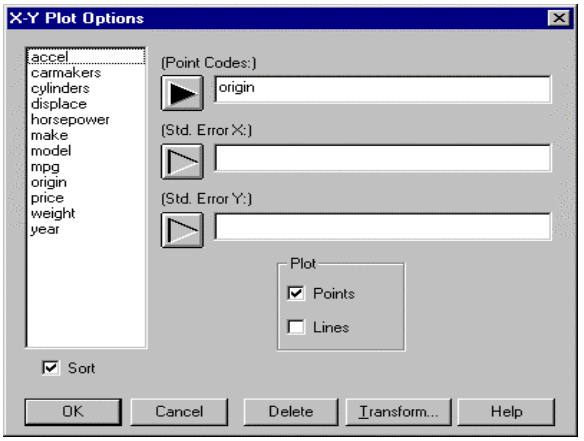


Figura 26

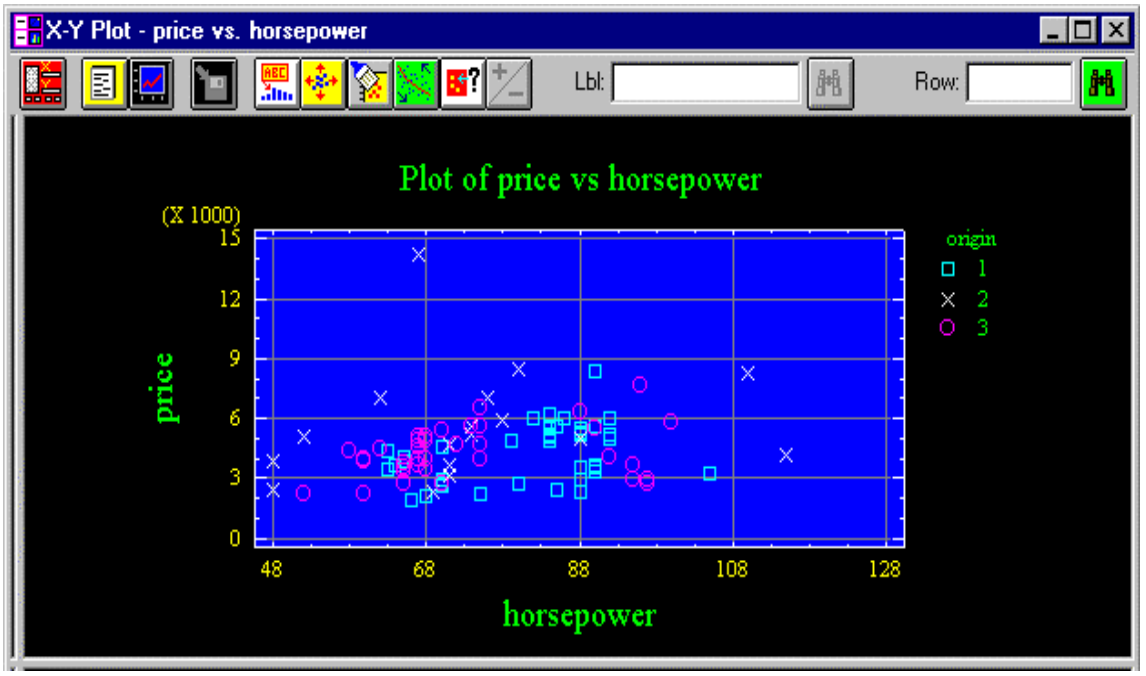


Figura 27

El procedimiento PLOT… SCATTERPLOTS… MULTIPLE X-Y PLOT permite superponer varios gráficos planos sobre el mismo par de ejes, dado que la variable X (abscisa) sea la misma en todos ellos. Es adecuado para comparar tendencias por grupos. El eje de ordenadas admite dos graduaciones (a la derecha e izquierda del gráfico).

Los procedimientos X-Y-Z PLOT y MULTIPLE X-Y-Z PLOT son totalmente análogos a los dos vistos anteriormente. La única diferencia es que incorporan una nueva variable para representar los puntos, por lo que los gráficos resultantes serán tridimensionales. Podemos ver un ejemplo en la *Figura 28*.

En dicho ejemplo hemos utilizado una nueva herramienta asociada a procedimientos gráficos. Picamos en la barra de herramientas de la ventana de análisis el icono *Identify* y nos aparecerá una lista en la que podemos seleccionar una variable por la que queremos identificar las observaciones. En nuestro caso hemos elegido MAKE. Así, al seleccionar con el ratón uno de los puntos del gráfico aparecerá su marca en el campo marcado como *Lbl:*

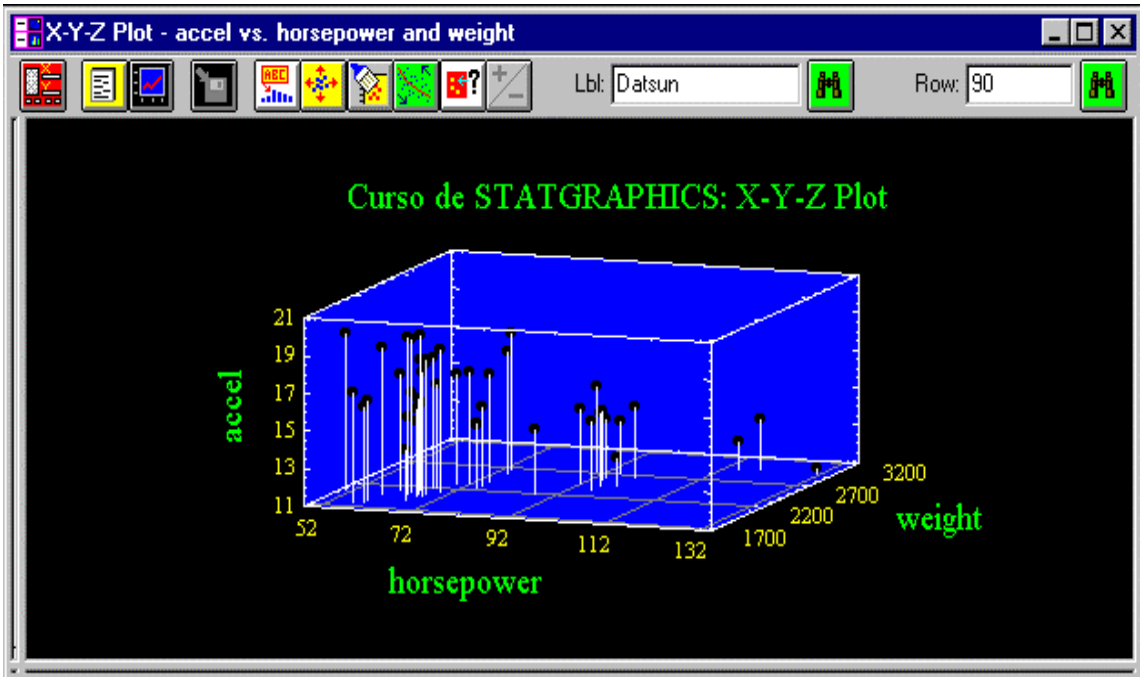


Figura 28

Gráficos de barras

Este tipo de gráfico es el más adecuado para tratar variables categóricas. Para cada una de las posibles categorías o valores, se dibuja una barra cuya longitud indica el número de observaciones con dicho valor para dicha variable. Podremos tratar una única variable o cruzar dos, eligiendo entre distintos posibles formatos para los gráficos resultantes.

Si tenemos las observaciones originales y queremos que el sistema cuente las que hay en cada categoría utilizaremos el procedimiento DESCRIBE… CATEGORICAL DATA… TABULATION para una única variable. Si lo que queremos es cruzar dos variables usaremos CROSSTABULATION. En ambos casos seleccionaremos BARCHART en *Graphics Options*. No podemos aplicar este procedimiento a variables que tengan excesivas categorías (el límite del sistema es 20), pues el gráfico resultante sería ilegible. Como ejemplo, vamos a trabajar con las variables ORIGIN y YEAR del fichero CARDATA (el caso de una única variable es muy sencillo y totalmente similar a los histogramas de frecuencia para variables numéricas que vimos anteriormente). Ya conocemos la primera y sabemos que tiene 3 categorías. YEAR contiene el año del modelo de automóvil y tiene 5 categorías: del 78 al 82.

En la pantalla de entrada de datos de CROSSTABULATION escribimos YEAR en *Row variable* (que será así tratada como variable primaria de clasificación) y ORIGIN en *Column Variable* (que será tratada como secundaria). Al seleccionar BARCHART en *Graphics Options* nos aparecerá el gráfico resultante. En *Pane Options* podremos seleccionar el tipo de gráfico, si queremos representar frecuencias absolutas o relativas y el tipo de orientación. En la *Figura 29* tenemos el gráfico en formato de barras múltiples (*clustered*, es el valor por defecto) con orientación vertical (el defecto es horizontal). En la *Figura 30* tenemos el mismo gráfico en formato de barras apiladas (*stacked*).

Si en lugar de tener el fichero con las observaciones individuales originales tuviéramos uno con una observación por cada categoría de la variable primaria y una o varias variables que contengan las frecuencias correspondientes a cada categoría, obtendríamos estos gráficos con los procedimientos PLOT… BUSINESS CHARTS… (MULTIPLE) BARCHART.

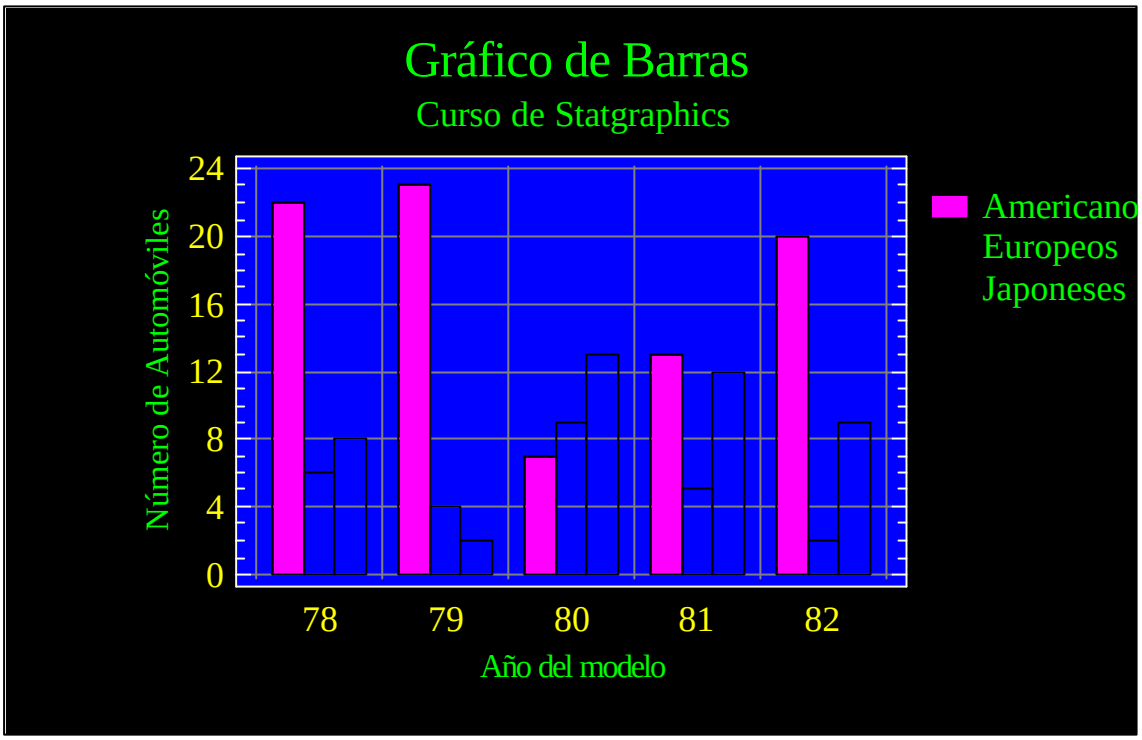


Figura 29

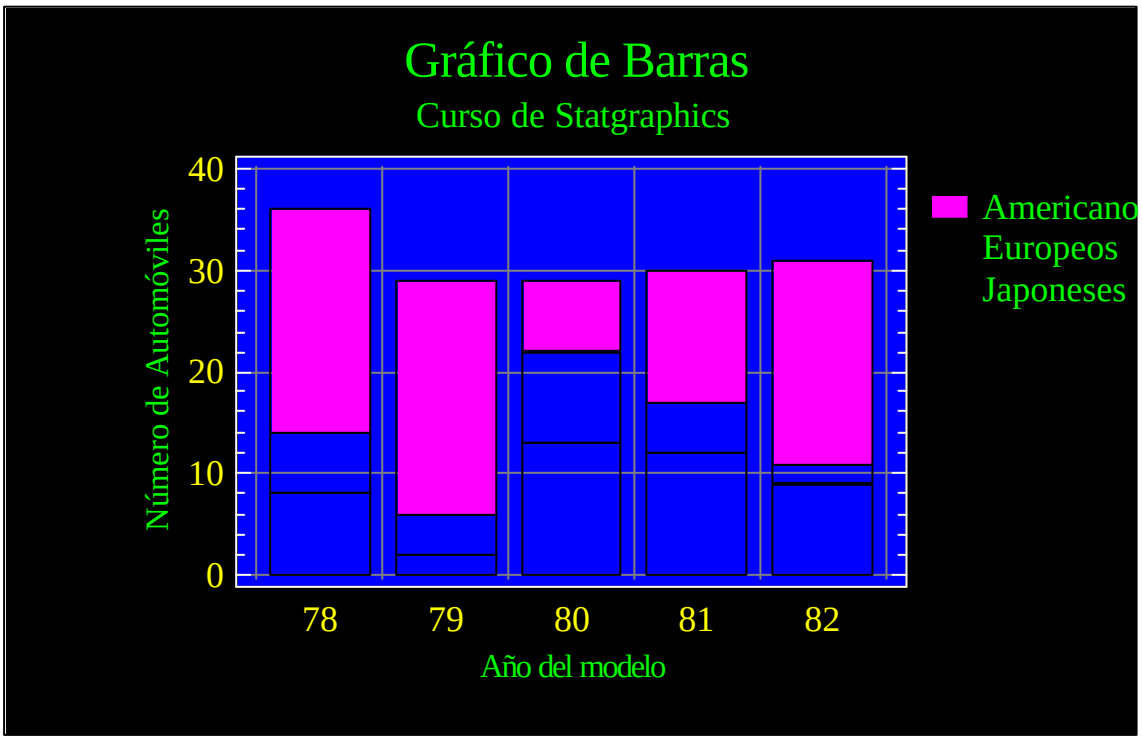


Figura 30

Gráficos de sectores

También idóneos para variables categóricas, representan los individuos (o porcentajes) en cada una de las categorías como sectores de un círculo. El círculo representa la totalidad (o el 100%) de las observaciones. Por su aspecto, estos gráficos reciben también el nombre de *gráficos de tartas.*

En este caso no se permite el cruce de variables, por lo que sólo se podrán representar los valores de una única variable en el mismo gráfico. Como en el caso anterior, si tenemos las observaciones individuales originales que el sistema debe agrupar entramos al procedimiento DESCRIBE… CATEGORICAL DATA… TABULATION. Seleccionamos entonces PIECHART en *Graphics Options*.

En la *Figura 31* vemos un ejemplo de este tipo de gráficos. Podemos ver en él la distribución por marcas (variable MAKE) de los modelos de automóviles japoneses de nuestro fichero CARDATA (para ello indicamos *origin=3* en el campo *Select:* de la pantalla de entrada de datos). El gráfico se ha personalizado cambiando alguna de las opciones por defecto en *Pane Options*. Así, el sector de la marca Mazda aparece separado del resto por haber indicado un 3 en el campo *Offset Slice:* (si dejamos este campo en blanco todos los sectores aparecerán juntos cerrando el círculo), y el tamaño del círculo corresponde a un *Diameter* del 70%.

Si lo que tenemos es una observación por categoría y una variable que contenga la frecuencia correspondiente a cada una de ellas, generaremos este gráfico con el procedimiento PLOT… BUSINESS CHARTS… PIECHART.

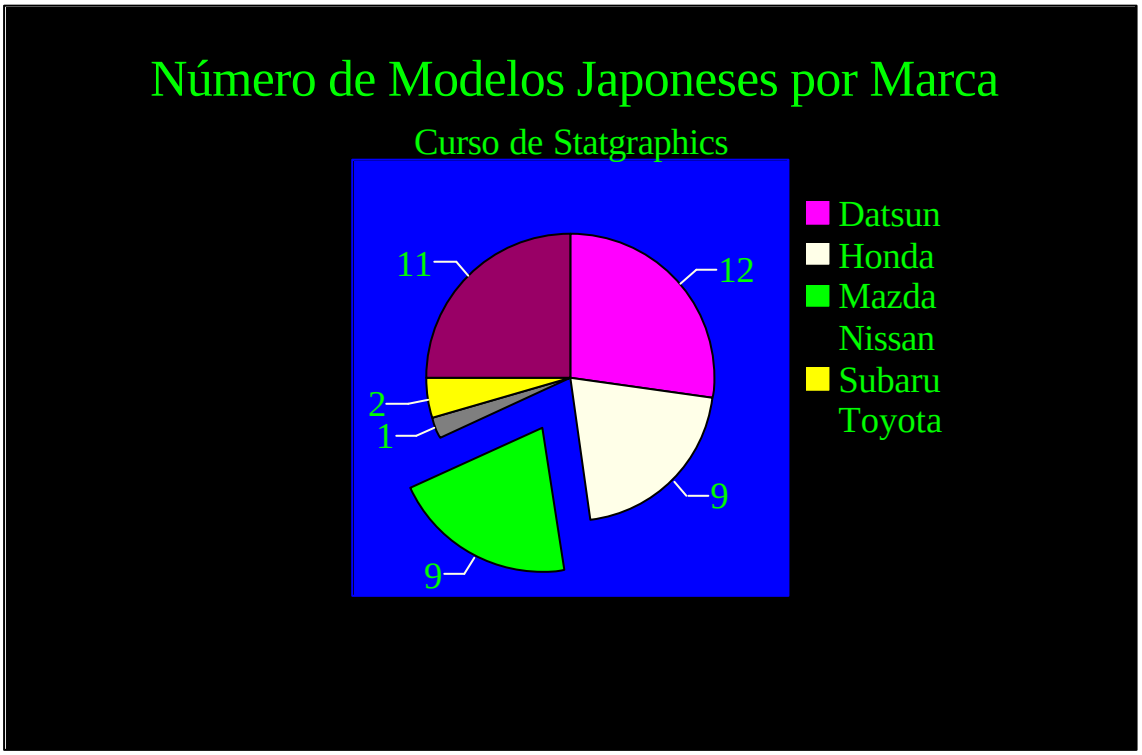


Figura 31

6. INFERENCIA ESTADÍSTICA

Conceptos

Hemos visto que nuestros ficheros de datos contienen información de ciertas características medidas en un determinado número de individuos. Sin embargo, no es corriente que nuestro objetivo sea extraer conclusiones sobre ese conjunto de individuos en particular.

Normalmente nuestro objetivo es más ambicioso. Queremos extender nuestras conclusiones a un conjunto más amplio de individuos que llamaremos *población* o *universo*. Dado que dicha población no es controlable exhaustivamente, recogemos la información en un subconjunto de individuos que consideramos suficientemente representativo de la misma, que llamaremos *muestra*. La *inferencia estadística* trata del conocimiento veraz de la población a partir del análisis de los datos muestrales.

Las características numéricas de las distribuciones poblacionales reciben el nombre de *parámetros*, mientras que sus equivalentes muestrales se llaman *estimadores*. A partir de los datos recogidos en nuestra muestra y almacenados como variables en nuestro fichero de datos, calcularemos los estimadores y aplicaremos los métodos de inferencia estadística para determinar qué grado de conocimiento nos aportan acerca de los auténticos parámetros poblacionales.

Realizaremos inferencias sobre la media y la varianza, referidas a una o dos muestras. Existen distintos métodos de estimación, de los que utilizaremos dos: los *intervalos de confianza* y los *contrastes de hipótesis*.

Veamos brevemente los conceptos que subyacen bajo ambos métodos. Un estimador nos proporciona una estimación puntual del parámetro, un único número que lleva asociada una medida de precisión llamada *error estándar* (el error estándar nos cuantifica la certidumbre que podemos tener sobre lo cerca que puede estar o no nuestra estimación del auténtico valor poblacional). El método de estimación por intervalos extiende estos conceptos a la generación de un intervalo que contendrá el auténtico valor del parámetro con un determinado nivel de confianza.

En cuanto al contraste de hipótesis, consiste en formular una conjetura previa acerca del parámetro (llamada *hipótesis nula* y representada habitualmente por H_0:) y ver si nuestros datos muestrales aportan evidencia que la soporte o, por el contrario, la contradiga. Para ello se calcula un estadístico (función de la muestra) cuya distribución sea conocida si la hipótesis nula es cierta. Si el valor obtenido es muy improbable bajo dicha distribución, rechazaremos la hipótesis nula. En caso contrario solamente podremos afirmar que no tenemos evidencia para rechazarla.

En este capítulo veremos cómo realizar inferencia basada en una o dos muestras. Para el primer caso usaremos el procedimiento DESCRIBE… NUMERIC DATA…ONE VARIABLE ANALISIS que ya conocemos. Para el segundo usaremos los procedimientos asociados a la opción COMPARE… TWO SAMPLES, cuyo menú asociado podemos ver en la *Figura 32*.

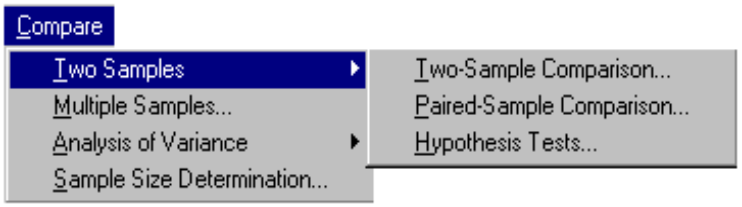


Figura 32

Inferencias basadas en una única muestra

Los análisis finales presentados como CONFIDENCE INTERVALS e HYPOTHESIS TESTS en *Tabular Options* del procedimiento DESCRIBE…NUMERIC DATA… ONE SAMPLE ANALISIS (ver *Figura 16*) realizan inferencias sobre la media y la varianza poblacionales de una variable a partir de los datos muestrales contenidos en un fichero de datos. El primero muestra el valor de los estimadores y construye intervalos de confianza. El segundo permite contrastar la hipótesis de que la media y la mediana tomen un valor concreto, frente a que tomen un valor mayor, menor o, simplemente, diferente (*hipótesis alternativa*).

En la pantalla de entrada de datos del procedimiento debemos indicar, como siempre, la variable que contiene los datos muestrales. Después marcamos CONFIDENCE INTERVALS e HYPOTHESIS TESTS en *Tabular Options*. Se generan entonces dos paneles con los resultados de ambos análisis, que comentamos por separado a continuación.

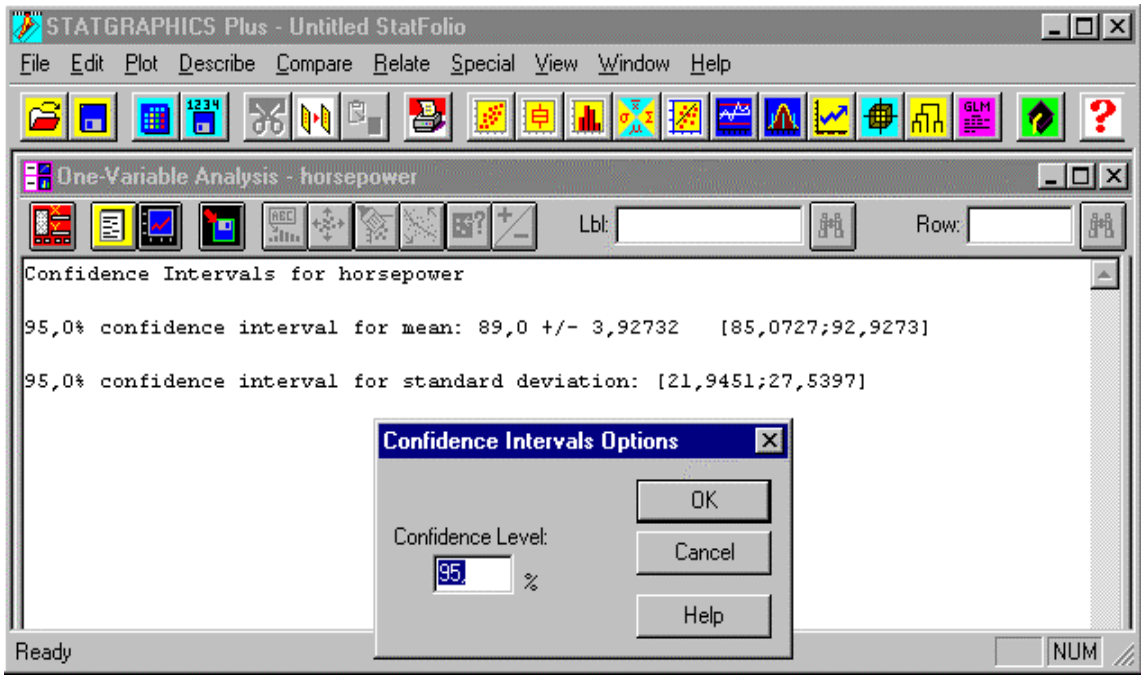


Figura 33

En nuestro ejemplo de la *Figura 33* hemos pedido el análisis de la variable HORSEPOWER. El intervalo de confianza creado para la media es (85.0727 , 92.9273), y para la desviación típica (21.9451 , 27.5397), ambos con 150 grados de libertad. La interpretación que debemos dar al nivel de confianza es que el intervalo construido de la misma manera para todas las posibles muestras del mismo tamaño, contendría al auténtico valor del parámetro el 95% de las ocasiones. En la figura también aparece la ventana asociada a *Pane Options* en la que sólo podemos modificar el nivel de confianza.

Por otro lado, con HYPOTHESIS TESTS queremos contrastar la hipótesis nula de que la auténticas media y mediana poblacionales sean 80. Indicaremos en la ventana de *Pane Options* los parámetros del modelo: el valor (80) a contrastar en la hipótesis nula (*Mean:*), la hipótesis alternativa (*Alt. Hypothesis*) y el *nivel de significación* que queremos fijar para el test (*Alpha:*). Debemos entender este último parámetro como el límite máximo que queremos permitir que tenga la probabilidad de equivocarnos en caso de rechazar finalmente la hipótesis nula. En la *Figura 34* podemos ver esta pantalla junto con los resultados del análisis. El estadístico computado en este test provendrá de una distribución *t-Student* si dicha hipótesis es cierta.

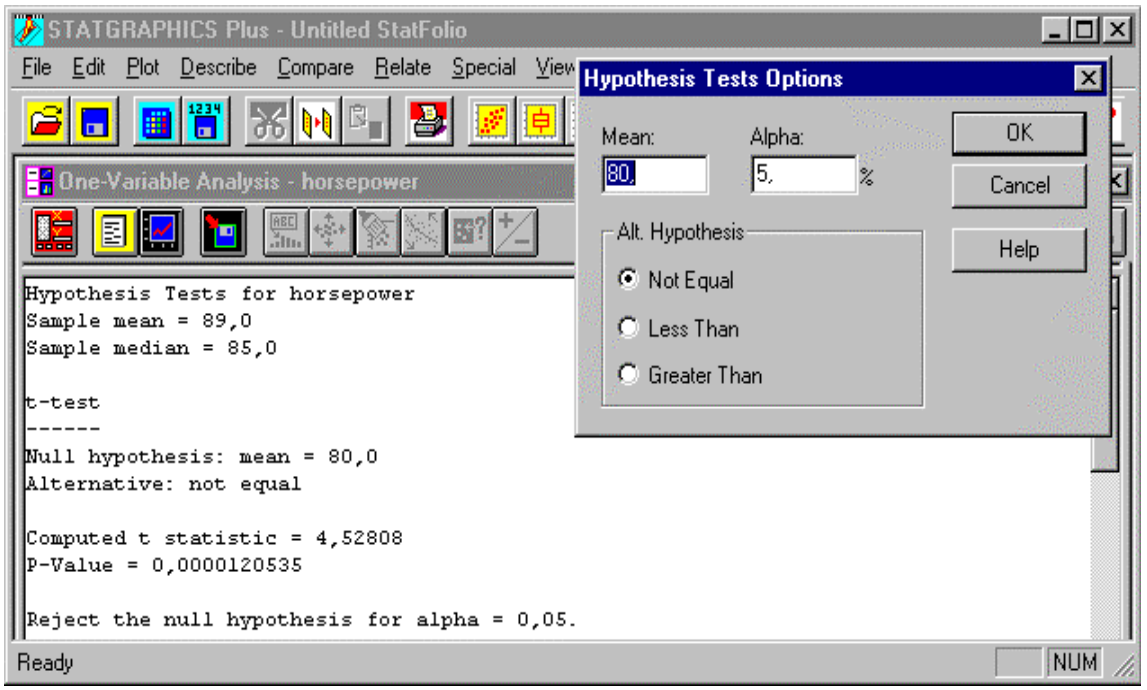


Figura 34

En nuestro caso, el valor resultante (*Computed t-statistic=4.52808*) es muy poco compatible con ella. De hecho, la probabilidad de obtener por azar un valor igual o mayor (en valor absoluto, pues hemos elegido como alternativa *Not Equal*) que 4.52808 bajo dicha distribución es sólo 0.0000120535 (*p-value*), mucho menor que el nivel de significación establecido por nosotros (*Alpha=0.05*). Por lo tanto, rechazamos la hipótesis nula (*reject the null hypothesis).* Esto es, nuestros datos no soportan la hipótesis de que el auténtico valor de la media poblacional sea 80.

Comparación de dos grupos (muestras independientes)

El procedimiento COMPARE… TWO SAMPLES… TWO-SAMPLE COMPARISON permite comparar los parámetros (media y varianza) de dos poblaciones a partir de muestras independientes.

Las dos poblaciones estarán definidas por una característica diferenciadora que permita clasificar a los individuos. Así, podríamos estar interesados en saber si la pérdida de función renal con la edad es más acusada o no en individuos de raza blanca y de raza negra, cuál de dos tipos de abono produce mayores rendimientos en una determinada explotación agraria, o si existen diferencias entre la capacidad pedagógica de dos diferentes métodos de aprendizaje.

Para realizar este tipo de análisis grabaremos los datos de la variable a analizar en un fichero, junto con otra que identifique a cuál de los dos grupos pertenece el individuo correspondiente. Así tendremos un fichero con, como mínimo, tantos registros como individuos en ambas muestras. Nuestro objetivo será comparar los parámetros poblacionales para la variable de análisis en los dos grupos.

Como ejemplo, compararemos la potencia de los coches americanos y europeos a partir de los datos muestrales contenidos en nuestro fichero CARDATA. Al invocar al procedimiento, en la pantalla de entrada de datos de la *Figura 35* indicaremos en el campo *Input:* que la estructura de datos se corresponde con la explicada en el párrafo anterior (*Data and code columns*). La otra opción (*Two data columns*) nos permitiría haber grabado los datos de cada

grupo como dos variables diferentes. La variable de análisis se identificará en el campo *Data:* y la que define los grupos en *Sample Code:* (HORSEPOWER y ORIGIN respectivamente en nuestro caso). En el campo *Select:* escribimos una proposición lógica que excluya los automóviles japoneses del análisis.

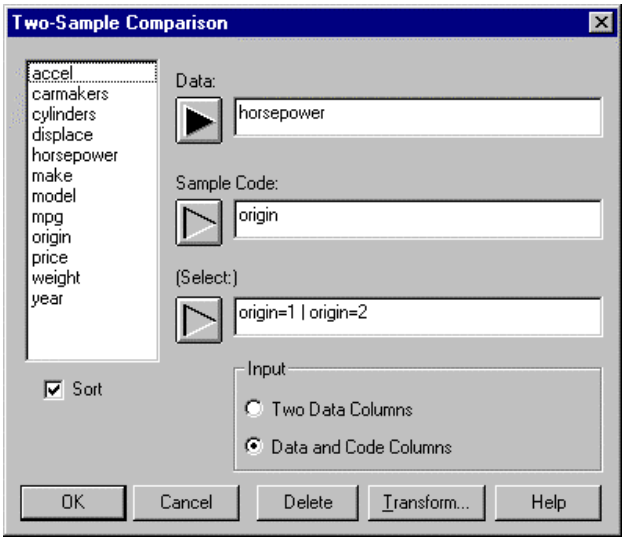


Figura 35

Seleccionamos después dos análisis finales en *Tabular Options*: *Comparison of Means* y *Comparison of Standard Deviations*.

El test utilizado para la comparación de medias se basa también en la distribución *t-Student*, pero el cálculo del estadístico asociado depende de si podemos considerar o no las varianzas (o desviaciones estándar) diferentes en ambos grupos. Por lo tanto, tendremos que interpretar primero la salida producida por COMPARISON OF STANDARD DEVIATIONS, que podemos observar en la *Figura 36*.

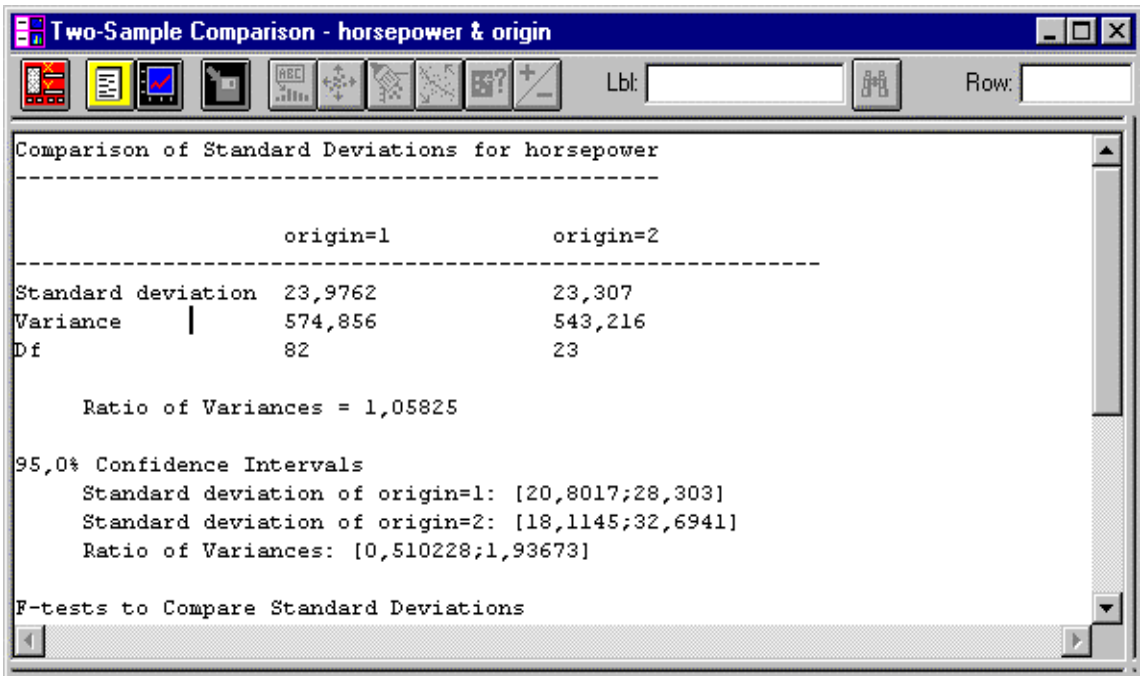


Figura 36

Debemos fijarnos en primer lugar en el intervalo de confianza para el cociente de varianzas (*Ratio of Variances*). Dado que, en nuestro caso, dicho intervalo contiene al 1, no podemos rechazar que ambas varianzas sean iguales. Un contraste de hipótesis posterior basado en la distribución *F de Fisher-Snedecor* (para el que la hipótesis nula sería "H_0: la desviación típica es igual en ambos grupos"), confirma este hecho (*p-value=0,916316*).

La comparación entre medias se realiza mediante inferencia sobre la diferencia entre ambas. El intervalo de confianza para la diferencia entre medias depende, como dijimos, de si

existe o no igualdad de varianzas. En la *Figura 37* podemos ver la salida asociada a COMPARISON OF MEANS.

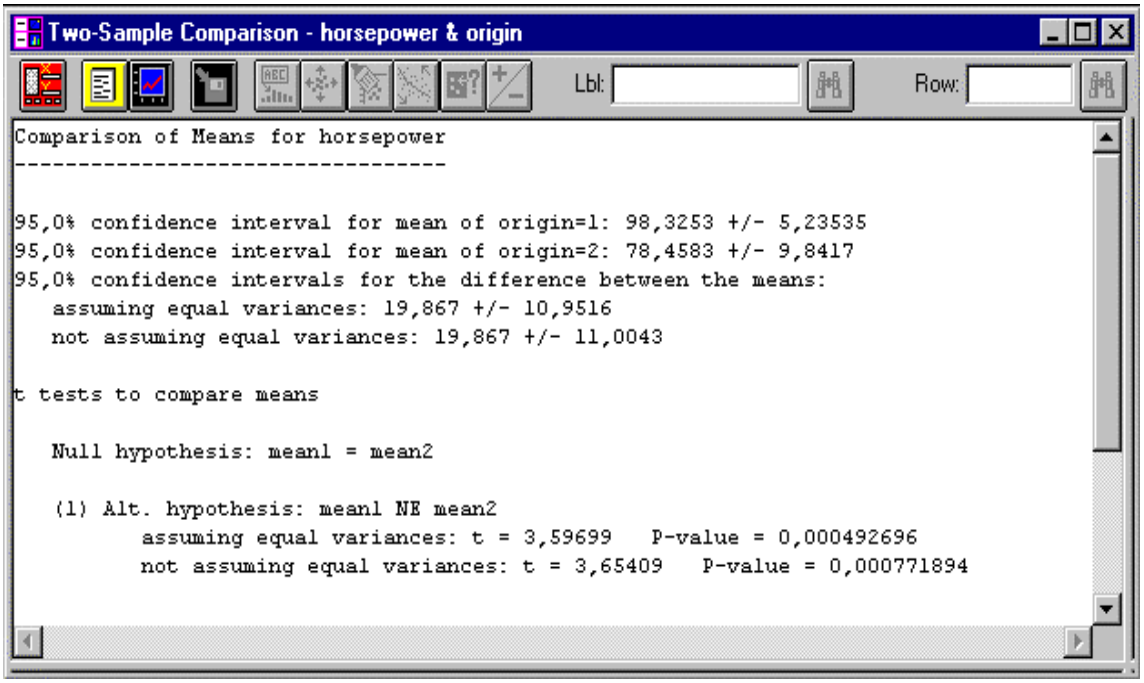


Figura 37

Dado el resultado obtenido para las desviaciones estándar, en nuestro caso debemos fijarnos en el intervalo generado para *assuming equal variances,* que es (8.9154,30.8186). Es especialmente reseñable el hecho de que el intervalo construido no contenga el valor 0.

Fijémonos ahora en el contraste H_o*:mean1=mean2* (también para el caso *assuming equal variances*). El valor del estadístico *t* obtenido (*3.59699*) no es compatible con la distribución teórica de la que debe provenir si la hipótesis nula es cierta (*p-value= 0.00049*). Por lo tanto, si fijamos un nivel de significación habitual como 0.05, debemos concluir que nuestros datos aportan evidencia de que, en media, los coches americanos son más potentes que los europeos.

Tanto en el caso de comparación de desviaciones típicas como de medias, podremos cambiar el nivel de confianza de los intervalos generados en *Pane Options*.

Si accedemos en este procedimiento a *Graphics Options*, se nos ofrecen varios procedimientos para comparar ambos grupos gráficamente. De entre ellos destacamos los llamados "gráficos de cajas y bigotes" (*box and whisker plots)* que vemos en la *Figura 38*.

Los gráficos de barras y bigotes son otra herramienta descriptiva que nos permite conocer las características generales de la distribución de una variable a partir de los datos muestrales. Consta de una caja central cuya amplitud es el *rango intercuartílico*, esto es, el 50% central de la distribución. La caja es atravesada por una línea que representa la *mediana*, que por lo tanto la divide en dos subcajas que contienen el mismo porcentaje de distribución (25%). El 50% restante de la distribución (25% por encima de la caja y 25% por debajo) se representa de la siguiente manera: como una línea continua el rango que contiene aquellos datos que no se separan del cuartil más cercano más de 1.5 veces el rango intercuartílico, y como puntos individuales el resto, que son así considerados como atípicos.

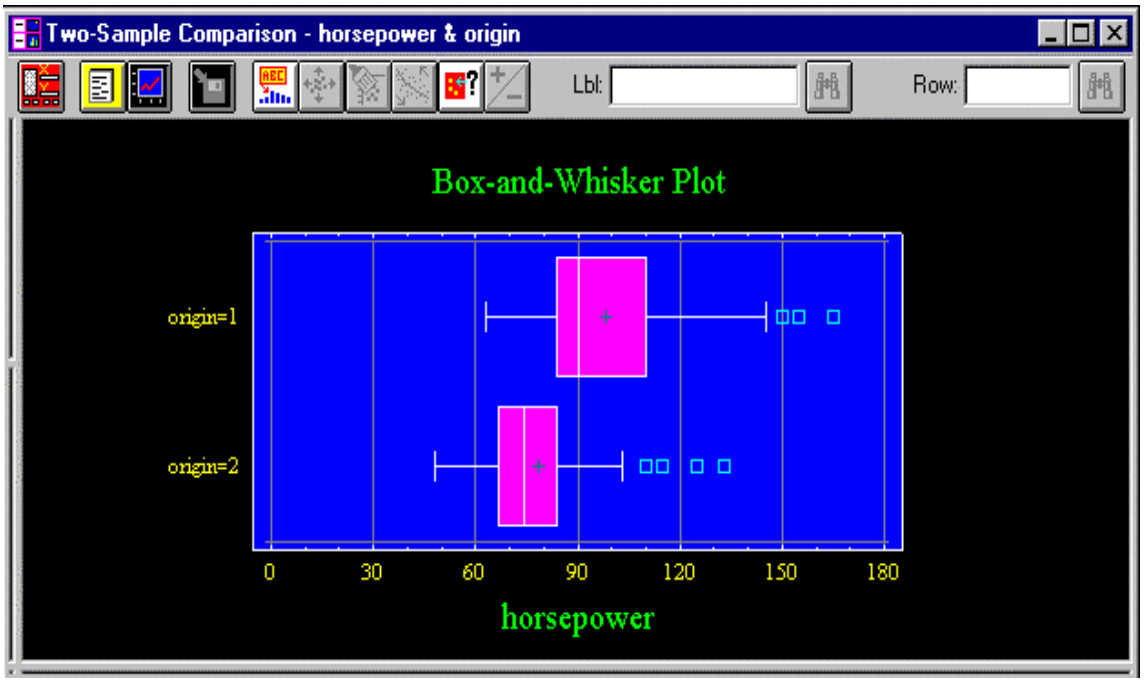


Figura 38

Estos gráficos son útiles para estudiar ajustes a distribuciones determinadas, simetrías y detección de valores atípicos. Son directamente accesibles desde los procedimientos PLOT… EXPLORATORY PLOTS… (MULTIPLE) BOX AND WHISKER PLOT, así como desde *Graphics Options* en el procedimiento DESCRIBE… NUMERIC DATA… ONE VARIABLE ANALYSIS.

Por último, es importante remarcar la importancia de la elección de la muestra que representará a cada grupo para obtener resultados fiables. Las muestras elegidas deben ser homogéneas respecto a otros criterios que puedan influir en la variable de análisis pues, si no, no podremos estar seguros de que las diferencias encontradas se deban realmente a la pertenencia a uno de los grupos. Si en el caso de la función renal nuestra muestra de individuos de una de las razas, digamos la blanca, esta compuesta por individuos más hipertensos o que consumen más alcohol en media que los que componen la muestra de individuos de raza negra, podríamos llegar a la conclusión de que las diferencias encontradas en la función renal de ambos grupos se deben a la raza. Sin embargo eran otros factores los que estaban influyendo en nuestra variable de análisis, y el sesgo en la muestra nos llevó a conclusiones erróneas.

En el caso anterior los grupos venían diferenciados por una característica inherente a los individuos. En otras ocasiones la definición de los grupos depende del experimentador. Así, en el ejemplo de los abonos seleccionamos un serie de parcelas para realizar un mismo tipo de cultivo. Podemos decidir en qué parcelas aplicamos un tipo de abono y en cuáles otro. Es obvio que la calidad del suelo de las parcelas influirá en el rendimiento, y que si el reparto no es homogéneo podemos achacar al tipo de abono diferencias que no se deben a él. Lo mismo ocurre en la evaluación de dos métodos de aprendizaje. Si el grupo de niños al que se aplica el primer método es más inteligente en media que el grupo al que se aplica el segundo, llegaremos a una evaluación errónea de ambos métodos.

Debemos ver siempre en la *aleatorización* el único método fiable para prevenir la aparición de estos sesgos sistemáticos que pueden influir negativamente en el resultado de nuestros análisis.

Comparación de dos grupos (datos pareados)

Veremos ahora una técnica que evita en gran parte la aparición de sesgos como los comentados en el apartado anterior. Recibe el nombre de *muestreo por pares* o *comparaciones pareadas*, y consiste en la elección de la muestra por pares cuyos miembros sean lo más homogéneos posible para todos los aspectos que puedan influir en la variable de análisis. Después se asigna un tratamiento a una de las unidades experimentales en cada par y el otro tratamiento a la otra (también es conveniente que esta asignación sea aleatoria en cada par). Si los pares eran realmente homogéneos, las diferencias encontradas al comparar ambos grupos pueden ser realmente achacadas al tratamiento recibido.

Veamos la diferencia entre este enfoque y el anterior con el ejemplo de los abonos. Supongamos que hemos elegido diez parcelas separadas para realizar los cultivos, de manera que cada parcela es homogénea en cuanto a composición, condiciones climatológicas y, en definitiva, cualquier criterio que pueda influir en el rendimiento de la misma. En el enfoque anterior (muestras independientes) asignaríamos aleatoriamente cinco parcelas para ser tratadas con un tipo de abono y las otras cinco con el otro. Con el enfoque de datos pareados dividiríamos cada una de las diez parcelas en dos mitades, asignando de forma aleatoria el abono correspondiente a cada mitad.

Un caso particular es cuando el par de mediciones proviene de la misma unidad experimental. Así, si queremos evaluar la eficacia de un determinado tratamiento, podemos medir las variables pertinentes en los mismos individuos antes y después de recibirlo. Este tipo de estudios reciben el nombre de *medidas repetidas*.

Para tratar los datos pareados con STATGRAPHICS no podemos utilizar el procedimiento del apartado anterior, puesto que el modelo que utiliza sólo es válido para muestras independientes y , claramente, nuestro nuevo enfoque viola esta hipótesis.

El proceso que debemos seguir es el siguiente. Almacenaremos nuestros datos en un fichero STATGRAPHICS con la siguiente estructura: un único registro para cada par, y dos variables que contengan los valores correspondientes a cada uno de los tratamientos (grupos). Después invocamos el procedimiento COMPARE… TWO SAMPLES… PAIRED SAMPLE COMPARISON e indicamos los nombres de las variables en los campos *Sample1* y *Sample2* de la pantalla de entrada de datos que se nos presenta.

Como ejemplo, supongamos que queremos contrastar la eficacia de un nuevo medicamento contra la hipertensión. Los datos tensionales de una muestra de siete pacientes antes de serles suministrado el mismo son 98.2, 107.9, 95.3, 102.1, 99.9, 98.7 y 100.5. Después del tratamiento las cifras fueron respectivamente 98.1, 103.0, 96.5, 98.3, 89.8, 90.6 y 94.1. Grabamos pues estos datos en un nuevo fichero en dos variables llamadas, por ejemplo, PRE y POST.

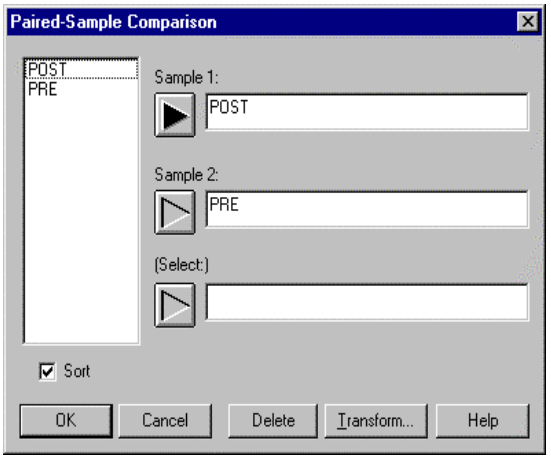


Figura 39

Indicaremos POST como *Sample1* y PRE como *Sample2* en la pantalla de entrada de datos como se muestra en la *Figura 39*. En realidad, el procedimiento va a realizar inferencia sobre la media de la variable diferencia POST-PRE. Dado que la media de esta nueva variable es la diferencia de las medias de las variables originales, nos interesa contrastar la

hipótesis nula de que su auténtico valor poblacional sea cero. Si podemos rechazar dicha hipótesis, concluiremos que ha habido efecto del tratamiento (las medias de PRE y POST serían entonces significativamente diferentes).

Para realizar el contraste seleccionamos HYPOTHESIS TESTS en *Tabular Options*. Los resultados se muestran en la *Figura 40*, y nos llevan a aceptar la efectividad del medicamento (la media resultante es negativa y el contraste permite rechazar que no haya diferencias).

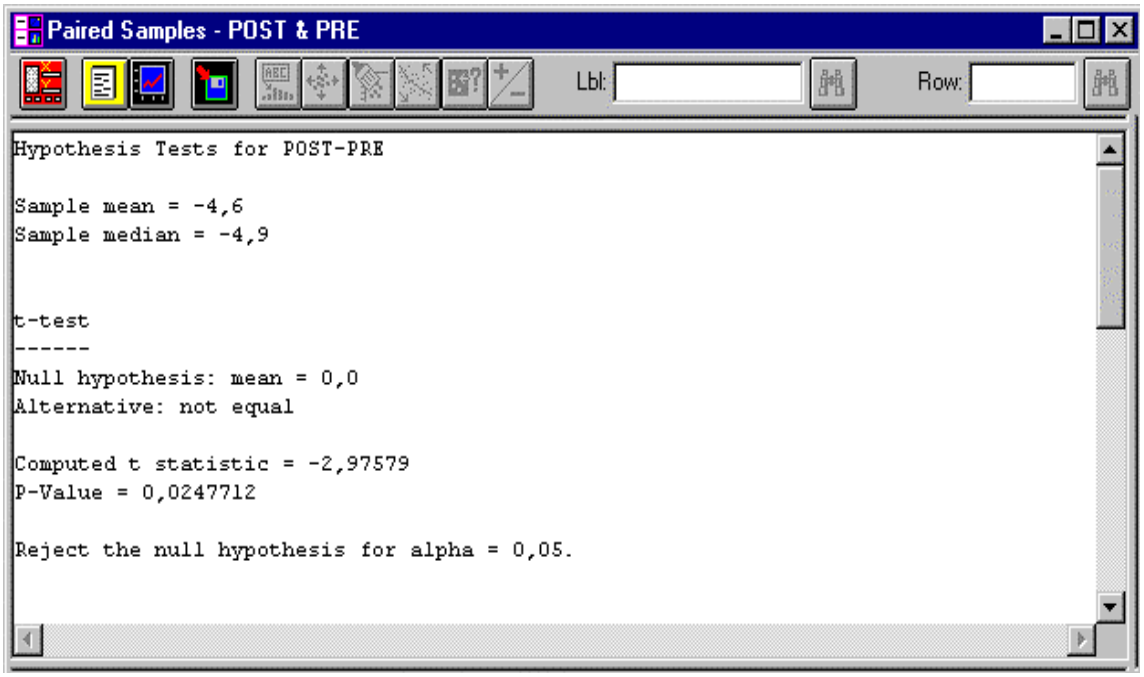


Figura 40

La hipótesis de normalidad

Aunque no lo hemos mencionado anteriormente, los métodos de inferencia expuestos (a excepción de los basados en las medianas) requieren, cuando la muestra es pequeña, que las variables provengan de la *distribución normal* (esto es, que se ajusten a la famosa *campana de Gauss*).

Los métodos de inferencia que requieren una estructura específica en las distribuciones poblacionales reciben el nombre de *paramétricos*, y nos plantean dos problemas fundamentales: cómo saber si los datos provienen de una distribución específica (en nuestro caso, la normal), y qué hacer cuando esto no ocurre.

En cuanto a la primera cuestión la opción NORMAL PROBABILITY PLOT, en *Graphics Options* del procedimiento DESCRIBE… NUMERIC DATA… ONE VARIABLE ANALYSIS, nos permite una evaluación visual del ajuste de nuestra variable a la distribución normal. El procedimiento sólo requiere como entrada el nombre de la variable y nos presenta un gráfico bidimensional como el de la *Figura 41* (para la variable HORSEPOWER). La línea recta representa la distribución normal teórica y los puntos son los datos de nuestra variable. Si nuestros datos provienen de una normal se ajustarán bastante a la recta. En el ejemplo, vemos una importante desviación de la normalidad en la variable HORSEPOWER, que nos haría recapacitar sobre los resultados obtenidos anteriormente (también podríamos usar para una evaluación visual los histogramas que vimos en el Capítulo 4).

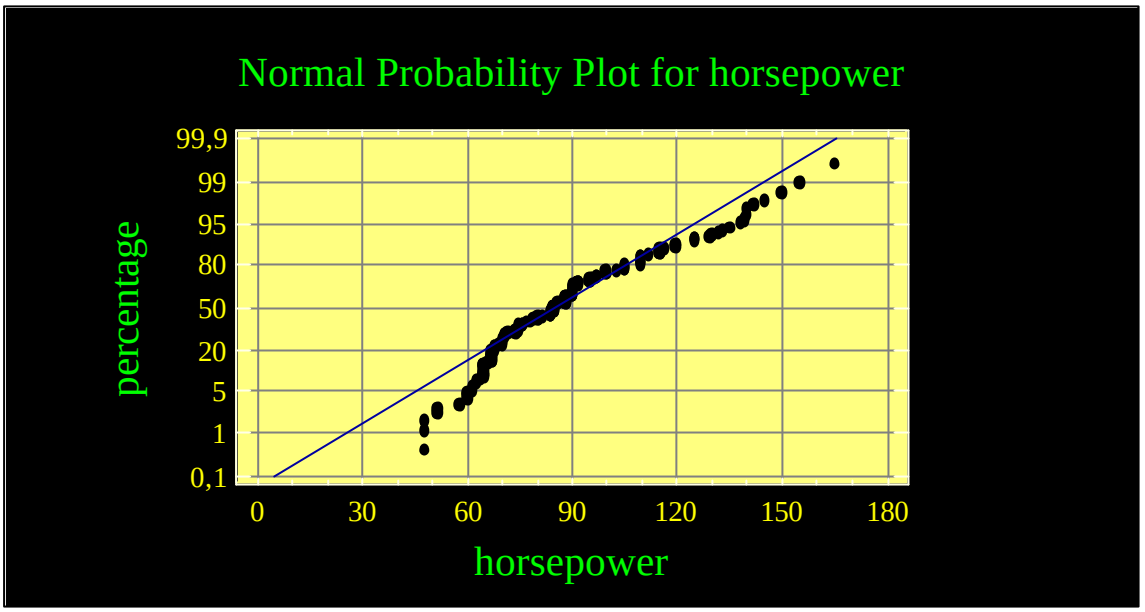


Figura 41

Existen contrastes para evaluar de manera objetiva la hipótesis de normalidad. Entre ellos está el de Kolmogorov-Smirnov, que permite contrastar el ajuste a cualquier distribución. Junto con otros, está accesible desde el procedimiento DESCRIBE… DISTRIBUTIONS... DISTRIBUTION FITTING(UNCENSORED DATA)

En cuanto al segundo problema planteado, qué hacer si los datos no son normales, tenemos dos posibles soluciones. La primera consiste en aplicar transformaciones que conviertan las variables en normales y trabajar con estas variables transformadas. El problema será entonces interpretar los resultados en función de las variables originales.

La otra alternativa es trabajar con métodos que no requieran ninguna suposición sobre la distribución de los datos. Estos métodos reciben el nombre genérico de *no paramétricos*, y se basan en ciertas características de los datos como el signo o el rango. Para la comparación de grupos son famosos el Test de Signos, el Test de Rangos Signados de Wilcoxon y el Test de Suma de Rangos de Wilcoxon (equivalente al Test de Mann-Whitney). Los dos primeros son para datos pareados y sus resultados aparecen junto al *t-test* que vimos en el apartado anterior en el procedimiento PAIRED SAMPLE COMPARISON. El tercero es para grupos independientes, y se accede desde el procedimiento TWO SAMPLE COMPARISON seleccionando COMPARISON OF MEDIANS en *Tabular Options*.

7. ANÁLISIS DE LA VARIANZA

Conceptos

Supongamos que observamos una característica continua en una muestra de una población, y consideramos una o varias características cualitativas que dividen la población (y la muestra) en grupos. Nuestro objetivo será inferir si estas últimas nos aportan información sobre la variable continua observada. Esto es, queremos saber si nuestros datos aportan evidencia de que el valor esperado de nuestra variable continua es diferente en las subpoblaciones o grupos definidos por la(s) variable(s) categórica(s).

Ilustraremos la idea con un par de ejemplos:

a) En una explotación ganadera se dividen las reses en tres grupos, cada uno de los cuales se alimenta durante un tiempo con un tipo de pienso diferente. Posteriormente se mide el incremento promedio de peso de cada uno de los grupos. ¿Influye el tipo de pienso recibido en dicho incremento? De ser así, ¿qué tipo de pienso es el que proporciona una mayor ganancia de peso?

b) Se mide el rendimiento de un proceso químico realizado en distintas condiciones de luminosidad: alta, media y baja. En cada una de las condiciones se prueban además dos reactivos diferentes. ¿Influyen las condiciones de luminosidad en el rendimiento? ¿Influye el reactivo utilizado? Más aún, ¿La influencia de las condiciones de luminosidad se manifiesta de igual manera para ambos reactivos?

La variable continua de análisis recibe el nombre de *variable dependiente*. Las categóricas que definen los grupos suelen llamarse *variables independientes o explicativas* o, más comúnmente, *factores*. Los posibles valores que puede tomar un factor reciben el nombre de *niveles*, y cada posible combinación de los niveles de los distintos factores en estudio se llama *tratamiento*. Al efecto que se produce cuando la influencia de un factor sobre la variable dependiente es diferente según los distintos niveles de otro factor se le conoce como *interacción* entre ambos factores.

Así, en el ejemplo a) tenemos un único factor con tres niveles, mientras que en el b) tenemos dos factores con tres y dos niveles respectivamente, que producen seis posibles tratamientos. Al tener dos factores en el segundo ejemplo, tiene sentido preguntarnos sobre la interacción entre ambos.

La comparación entre dos grupos que vimos en el capítulo anterior es un caso particular del modelo con un único factor (con sólo dos niveles). Las consideraciones que se hicieron entonces sobre homogeneidad entre grupos respecto a otras variables, así como la importancia de la aleatorización como método de asignación de tratamientos siguen teniendo absoluta validez.

El modelo ANOVA (ANalysis Of VAriance) descompone la variabilidad total de la variable dependiente en componentes independientes que pueden ser atribuidas a distintas causas (factores e interacciones). El diseño del experimento en cuestión determinará el modelo matemático y la descomposición de la varianza a aplicar. No es objeto de esta documentación la exposición exhaustiva de los modelos matemáticos tratados.

STATGRAPHICS agrupa los procedimientos de Análisis de la Varianza bajo la opción COMPARE…ANALYSIS OF VARIANCE de la barra de menú. Podemos ver el submenú asociado en la *Figura 42*.

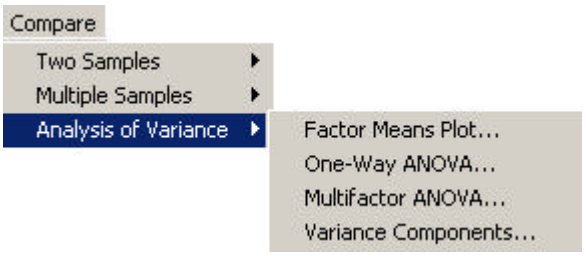


Figura 42

Abordaremos el modelo ANOVA con uno y dos factores, correspondientes a las opciones segunda y tercera del submenú.

Los modelos que veremos exigen fuertes restricciones sobre los datos: normalidad e igualdad de varianzas entre grupos (homocedasticidad) para la variable dependiente. Cuando estas restricciones sean violadas deberemos utilizar tests no paramétricos como el de Kruskal-Wallis, accesible desde *Tabular Options* del procedimiento COMPARE… ANALYSIS OF VARIANCE… ONE-WAY ANOVA. Como ya comentamos en el capítulo anterior, otra solución es aplicar transformaciones (normalmente dentro de la familia de Box-Cox) que generen variables que cumplan las restricciones, con el problema de interpretar los resultados en función de las variables originales.

Modelo con un factor

Queremos comparar la efectividad de tres analgésicos (etiquetados como 1, 2 y 3) en términos de la duración de su efecto en una muestra de 48 pacientes homogéneos a cada uno de los cuales se le suministra solamente uno de los mismos.

La manera en la que deben estar almacenados los datos para un análisis de este tipo es análoga a la que vimos para la comparación de dos grupos con muestras independientes: un registro por individuo, una variable (numérica o no) que contenga los niveles del factor y otra (numérica) con los valores de la variable dependiente.

En nuestro caso los niveles del factor correspondientes a cada individuo están almacenados en la variable PAINKILLER del fichero ANOVA, mientras que los valores de la variable dependiente están en la variable DURATION del mismo fichero.

Para llevar a cabo el análisis seleccionamos el procedimiento COMPARE… ANALYSIS OF VARIANCE…ONE-WAY ANOVA y nos aparecerá una sencilla pantalla de entrada de datos como la de la *Figura 43* donde debemos indicar los nombres de la variable dependiente y del factor.

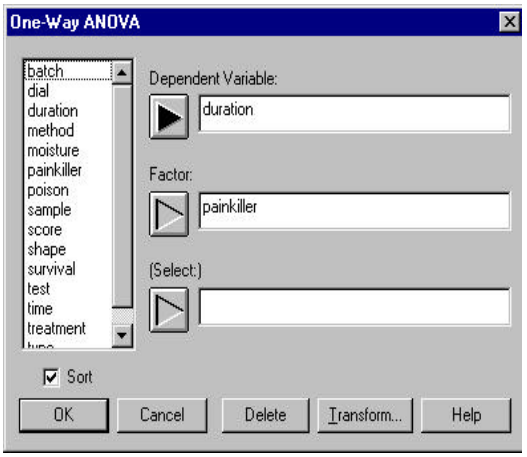


Figura 43

Seleccionaremos entonces ANOVA TABLE en *Tabular Options* y nos aparecerá como resultado una tabla conocida como Tabla del Análisis de la Varianza, de la que ya podemos extraer las primeras conclusiones (*Figura 44*).

Para ello, debemos saber que el modelo plantea un contraste de hipótesis para el cual la hipótesis nula es la igualdad de medias entre las subpoblaciones que definen los distintos niveles

del factor. En nuestro caso sería *H_0: El efecto de los tres analgésicos tiene la misma duración.* Se calcula entonces el valor de un estadístico (*F-ratio*) cuyo valor, de ser cierta la hipótesis nula, se sabe que proviene de una distribución F de Fisher-Snedecor. La probabilidad de obtener bajo esa distribución un valor igual o mayor que el resultante aparece en la tabla como *p-value*.

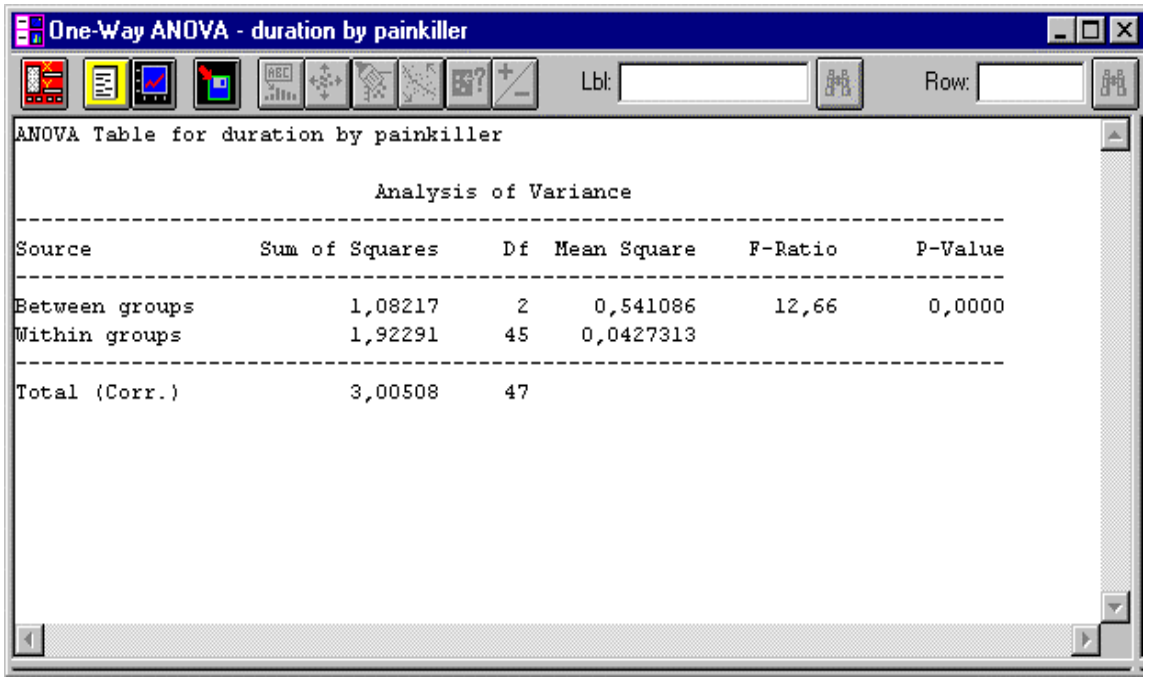


Figura 44

Si el valor obtenido es compatible con la hipótesis nula (*p-value* > 0.05) concluiremos que nuestros datos no aportan evidencia de que las medias poblacionales sean diferentes y nuestro análisis habrá concluido.

En caso contrario rechazaremos la hipótesis nula y aceptaremos, por lo tanto, que hay diferencias estadísticamente significativas entre los grupos. El valor de *p-value* podrá ser interpretado en tal caso como la probabilidad de habernos equivocado. Es habitual utilizar como valor límite el 0.05 nombrado anteriormente, que nos proporciona un nivel de confianza del 95%, aunque se pueden usar valores más restrictivos (0.01 para el 99%) o menos (0.1 para el 90%).

En caso de obtener un valor significativo, como en nuestro ejemplo, surge una pregunta inmediata. Aceptamos que hay diferencias, pero ¿cuáles son esas diferencias?. ¿Son todos los grupos significativamente diferentes entre si, o sólo alguno(s) de ellos con respecto a los demás?.

Para responder a estas preguntas necesitamos realizar un *Test de Rangos Múltiples* que compare los grupos por pares. Lo haremos seleccionando la opción MULTIPLE RANGE TESTS en *Tabular Options*. Para ver los distintos métodos de que dispone STATGRAPHICS para realizar este tipo de tests seleccionamos *Pane Options*: LSD (defecto), Tukey, Scheffe, Bonferroni, Newman-Keuls y Duncan (el elegido para nuestro ejemplo). El nivel de confianza se selecciona en el campo *Confidence Level* (95% por defecto).

Una vez seleccionado el método pulsamos OK y aparecerá el panel con los resultados. Los correspondientes a nuestro ejemplo se muestran en la *Figura 45*, que comentamos detalladamente a continuación.

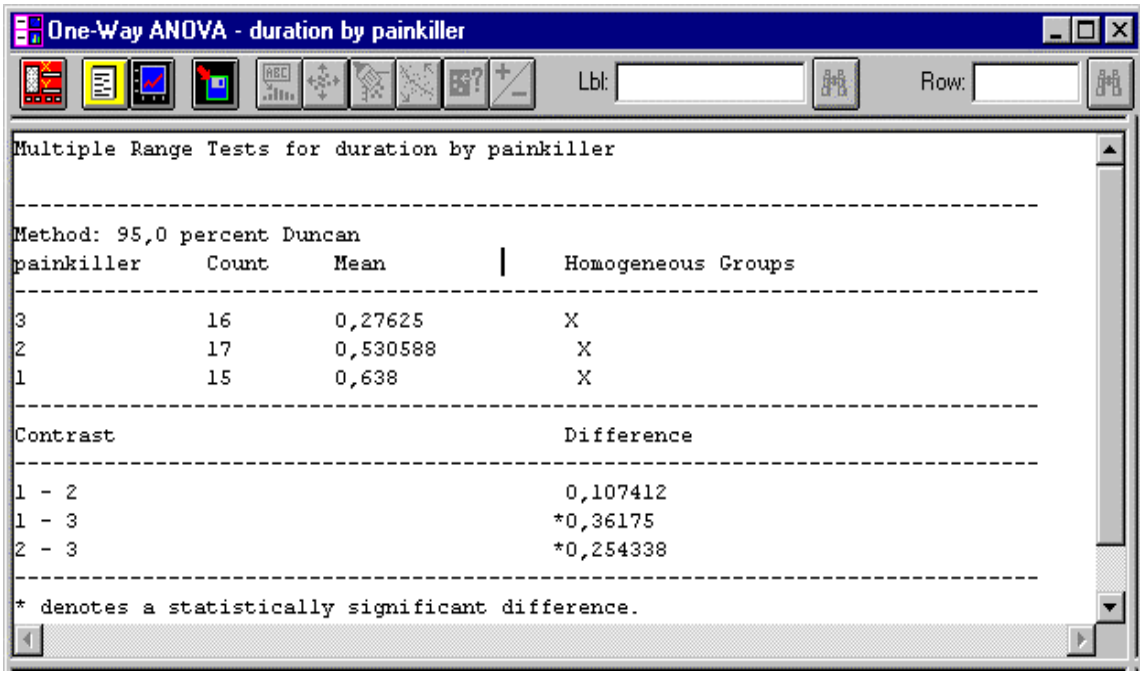


Figura 45

En primer lugar observamos una tabla con los distintos grupos o niveles que toma el factor, así como el tamaño de la muestra (*Count*) y la media muestral obtenida para la variable de análisis (*Mean*) en cada grupo. A la derecha aparece el campo *Homogeneous Groups*, que nos da ya las diferencias marcando cada grupo con uno o más asteriscos que deben interpretarse de la siguiente manera: consideraremos significativamente diferentes al nivel de confianza determinado las medias de aquellos grupos que no tengan ningún asterisco en la misma columna. No podremos concluir diferencias entre los grupos en caso contrario.

En nuestro ejemplo, no aparecen diferencias significativas entre la duración de los efectos de los analgésicos 1 y 2, mientras que ambos aparecen como significativamente mejores (el promedio de duración es mayor) que el analgésico 3.

La segunda tabla de la pantalla nos presenta la misma información con diferente formato. Para cada posible par de grupos (*Contrast*) nos aparece la diferencia entre sus medias muestrales (*Difference*) y un asterisco si ésta es estadísticamente significativa.

Comentamos a continuación las más interesantes del resto de opciones ofrecidas en *Tabular* y *Graphics Options*:

* *Means Table* produce una tabla con la estimación, error estándar e intervalos de confianza para la media en cada uno de los grupos.

* *Means Plot* presenta estos datos gráficamente y *Multiple Boxplot* presenta conjuntamente los "gráficos de cajas y bigotes" para cada grupo.

* *Residuals versus predicted*. Los *residuos* son las desviaciones cada valor observado para la variable dependiente a la media de su grupo. Estos gráficos son útiles para comprobar visualmente la hipótesis de homocedasticidad.

* *Variance Check* realiza los tests de Cochran, Bartlett, Hartley y Levene para comprobar dicha hipótesis objetivamente.

Modelo con dos factores

Si lo que queremos es estudiar simultáneamente el efecto de dos o más factores sobre la variable dependiente, utilizaremos el procedimiento COMPARE… ANALYSIS OF VARIANCE… MULTIFACTOR ANOVA, que presenta una pantalla de entrada de datos como la de la *Figura 46*. Vemos que en el campo *Factors:* tenemos la posibilidad de indicar múltiples factores. Para ilustrar este modelo utilizaremos el ejemplo a) del apartado introductorio de este capítulo, en el que añadiremos un nuevo factor.

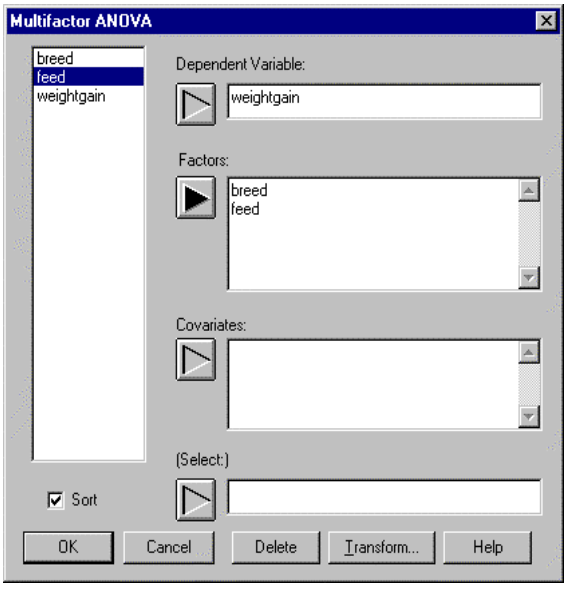


Figura 46

Queríamos estudiar el incremento de peso producido en ganado vacuno por tres tipos de pienso, que llamaremos A, B, y C. Nos planteamos ahora además si la raza del ganado puede tener efecto en dicho incremento. Los datos se encuentran en el fichero AG almacenados en tres variables, que son: FEED (tipo de pienso: A, B o C), BREED (raza: Angus, Charolais o Hereford) y WEIGHTGAIN (incremento de peso, nuestra variable dependiente).

Una vez declaradas las variables, seleccionamos ANOVA TABLE en *Tabular Options*. Por último, seleccionamos *Analysis Options* y escribimos un "2" en el campo *Maximum Order Interaction* que, por defecto, contiene el valor "1" (explicaremos más adelante el significado de este campo). Los resultados se muestran en la *Figura 46*. La tabla resultante tiene una estructura similar a la vista para un único factor, aunque ahora la parte de la variabilidad explicada por el modelo aparece a su vez descompuesta para poder contrastar cada uno de los posibles efectos.

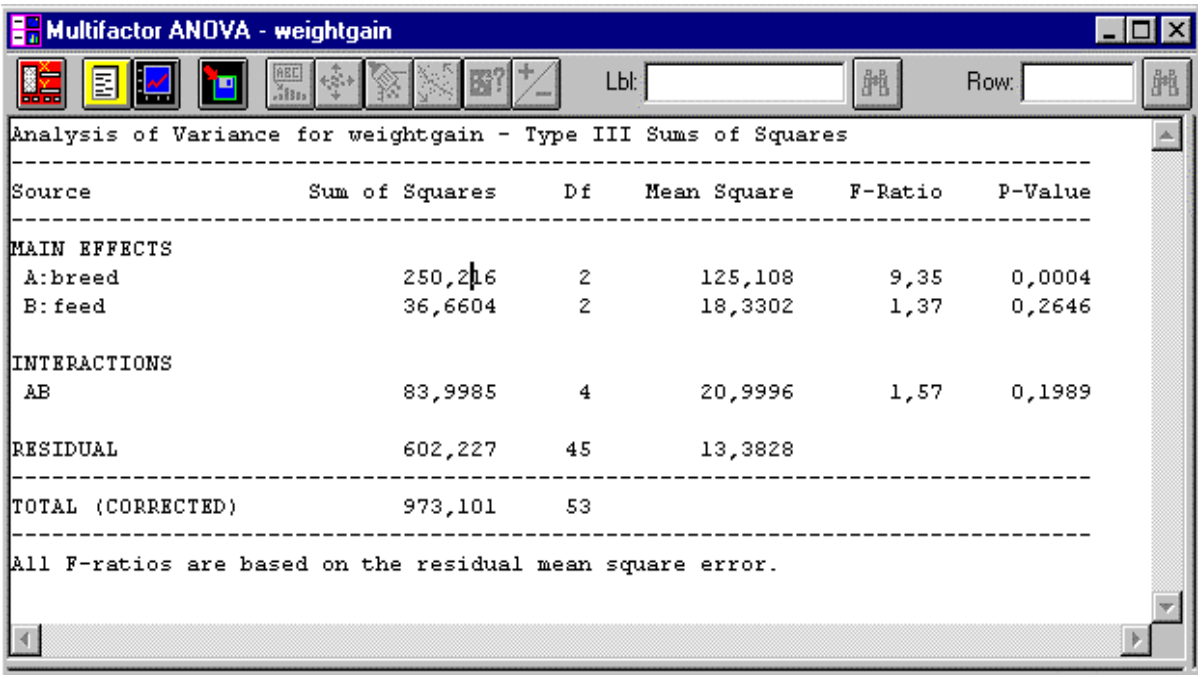


Figura 47

Nos fijamos ahora en los resultados obtenidos para los dos factores incluidos en el modelo (*MAIN EFFECTS*). Los niveles de significación del estadístico F son 0.2646 y 0.0004 para FEED y BREED respectivamente. Por lo tanto concluimos que el tipo de pienso no es significativamente influyente en el incremento de peso, mientras que las diferencias entre razas sí que aparecen como significativas.

Igual que en el caso de un único factor, queremos saber cuáles son las diferencias concretas entre los niveles de dicho factor. Como en el caso de un único factor seleccionamos la opción MULTIPLE RANGE TESTS en *Tabular Options*. Las tablas resultantes son exactamente iguales que en el caso unifactorial, pues el test se lleva a cabo factor a factor (en nuestro caso, se concluye que la raza Charolais es más productiva en términos de incremento de peso que las otras dos, que aparecen como no significativamente diferentes entre sí).

Interacción entre factores

Ya hemos definido anteriormente el concepto de interacción. En nuestro ejemplo su existencia supondría que las diferencias entre medias debidas al tipo de pienso son distintas según las razas o viceversa.

Para incluir en el modelo las interacciones entre factores (como hemos hecho en nuestro ejemplo) usaremos el campo *Maximum Order Interactions* de la pantalla que aparece al seleccionar *Analysis Options* desde el panel que contiene la tabla del análisis de la varianza. Si dejamos el "1" que trae por defecto no se incluirá ninguna interacción en el modelo. En nuestro caso hemos indicado un "2", que significa que queremos incluir la interacción entre los dos factores (única posible). Si tuviéramos un modelo con tres factores podríamos considerar tres interacciones de segundo orden y una de tercero (difícil de interpretar). Para incluir esta última indicaríamos un "3" y, para excluirla, un "2". Además, cuando hay varias posibles interacciones del mismo orden se puede elegir cuáles incluir en el modelo y cuáles no. Para ello utilizaríamos el botón *Exclude* que aparece en la misma pantalla.

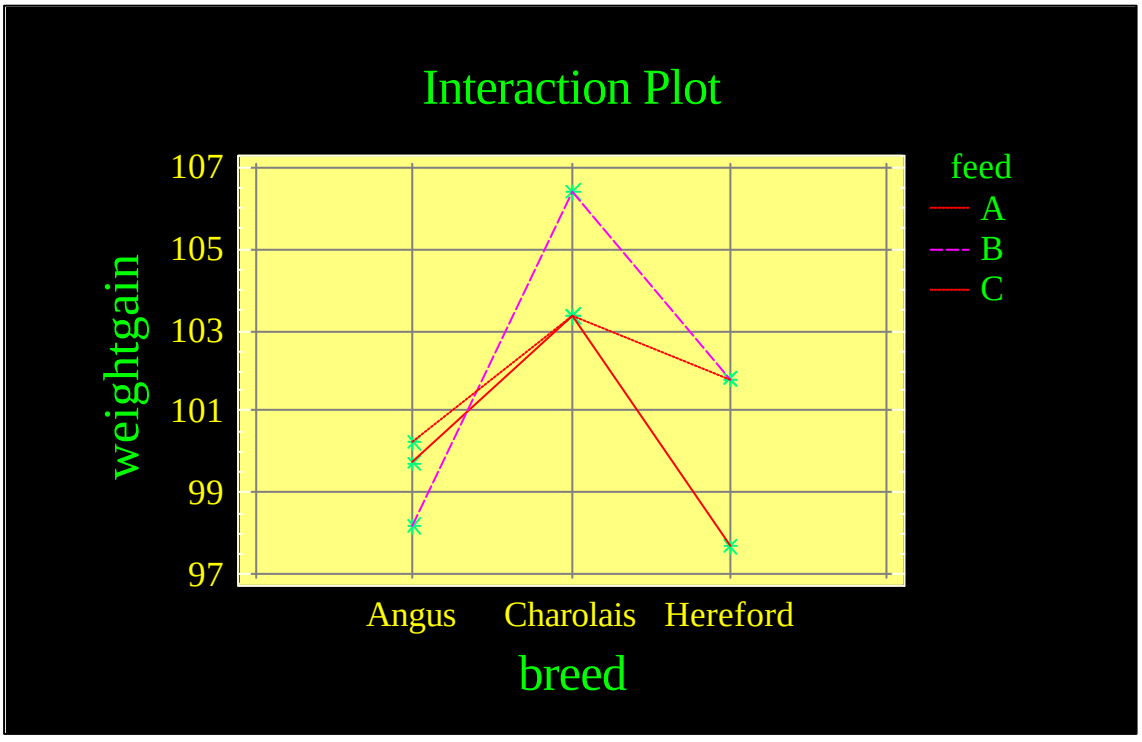


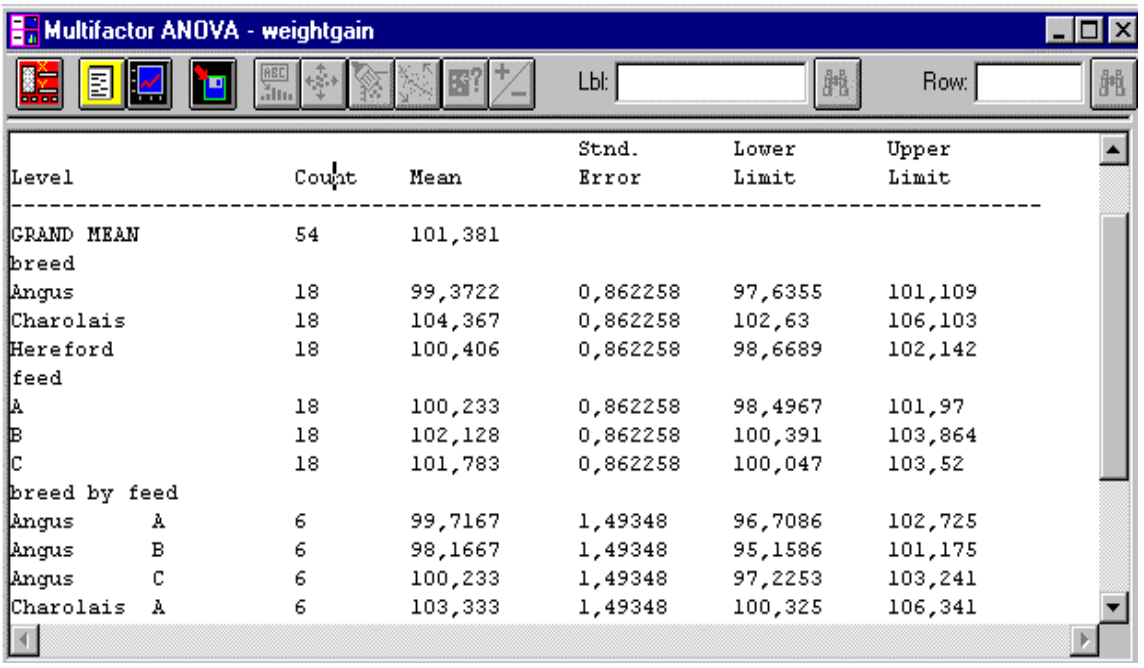
Figura 48

En la tabla de la *Figura 47* vemos la información referente a la interacción, separada de los efectos principales. La hipótesis nula será en este caso la ausencia de interacción. Dado el nivel de significación obtenido (0.1989) no podemos rechazar dicha hipótesis.

En el estudio de las interacciones son especialmente útiles los gráficos como el que se presenta en la *Figura 48*. Se construye desde *Graphics Options* seleccionando INTERACTION PLOT. Los niveles de uno de los factores se representan en el eje de abscisas, y para cada uno de ellos se representan las medias de la variable dependiente en cada uno de los niveles del otro factor. Después se unen con líneas los puntos correspondientes a medias provenientes del mismo nivel del segundo factor. Se observa así la evolución de las medias para el primer factor de manera independiente para cada uno de los niveles del segundo. Si las líneas resultantes tienen comportamientos muy diferentes indicará la existencia de interacción, concluyéndose ausencia de ésta cuando las líneas presenten comportamientos paralelos.

En nuestro ejemplo se observa un comportamiento similar para los tres tipos de pienso: subida en el incremento promedio de peso al pasar de Angus a Charolais, y luego bajada en los tres casos al observar los promedios en Hereford. Aunque las diferencias no eran las mismas para cada tipo de pienso (las líneas no son totalmente paralelas), no podemos concluir interacción significativa.

Para obtener una tabla con las medias e intervalos de confianza para cada nivel de todos los factores incluidos, así como de los cruces entre ellos (*Figura 49*) se elige la opción TABLE OF MEANS en *Tabular Options*. El resto de opciones gráficas o de texto para este procedimiento son similares a las explicadas en el caso unifactorial.



Level		Count	Mean	Stnd. Error	Lower Limit	Upper Limit
GRAND MEAN		54	101,381			
breed						
Angus		18	99,3722	0,862258	97,6355	101,109
Charolais		18	104,367	0,862258	102,63	106,103
Hereford		18	100,406	0,862258	98,6689	102,142
feed						
A		18	100,233	0,862258	98,4967	101,97
B		18	102,128	0,862258	100,391	103,864
C		18	101,783	0,862258	100,047	103,52
breed by feed						
Angus	A	6	99,7167	1,49348	96,7086	102,725
Angus	B	6	98,1667	1,49348	95,1586	101,175
Angus	C	6	100,233	1,49348	97,2253	103,241
Charolais	A	6	103,333	1,49348	100,325	106,341

Figura 49

8. REGRESIÓN LINEAL SIMPLE

Conceptos

Trataremos en este capítulo un problema similar al anterior. Queremos saber el conocimiento que una o varias *variables independientes* o *explicativas* nos aportan sobre el comportamiento de otra *variable dependiente* o *explicada*, que será una magnitud continua. La diferencia estriba en que, en este caso, las variables independientes serán también magnitudes *continuas*, en oposición a los factores categóricos que teníamos en el capítulo anterior.

Dispondremos de una muestra donde habremos observado tanto la variable dependiente como las independientes. La media muestral de la variable dependiente nos da una estimación general de la media poblacional. Intentaremos mejorarla estimando una media diferente para cada subpoblación definida a partir de unos valores concretos de las variables independientes.

Estas "medias condicionadas" serán función de los valores de las variables independientes, y la expresión explícita de la función que mejor ajuste el modelo dependerá de la naturaleza de la relación existente entre las variables. Esta es la idea de un modelo general de regresión. Nos restringiremos al caso en que dicha relación sea de tipo lineal (regresión lineal), y en el que sólo tengamos una variable independiente (regresión lineal simple).

En tal caso estimaremos dos parámetros, a (ordenada en el origen) y b (pendiente), que definen la recta con la que estimaremos el valor medio de la variable dependiente (y) a partir de cualquier valor particular de la variable independiente (x) de la forma:

$$y=a+bx$$

(Dicha recta será la que mejor se ajuste a la nube de puntos muestrales según el criterio conocido por el nombre de *mínimos cuadrados*). El valor así calculado podrá ser tomado también como predicción individual del valor de la variable dependiente, mejorando la estimación de la media muestral total si los datos realmente se ajustan al modelo lineal. Para los individuos de la muestra podemos calcular la desviación entre el valor observado y el predicho, que llamaremos *residuo*.

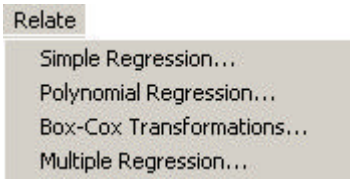


Figura 50

La *Figura 50* muestra el submenú asociado a la opción RELATE de la barra de menú, con los procedimientos de regresión que proporciona STATGRAPHICS.

El modelo de regresión lineal simple

Estudiaremos el modelo nombrado anteriormente, con una única variable independiente. Elegimos el procedimiento RELATE… SIMPLE REGRESION y nos aparecerá una pantalla de entrada de datos en la que sólo tendremos que indicar los nombres de las variables dependiente (en el campo *Y:*) e independiente (en el campo *X:*).

Utilizaremos de nuevo como ejemplo los datos de fichero CARDATA. Queremos ver en qué medida el conocimiento del peso de los automóviles nos aporta información sobre su potencia. Por lo tanto indicamos la variable HORSEPOWER como dependiente y la variable WEIGHT como independiente.

Al ejecutar el procedimiento los resultados se muestran en una pantalla como la de la *Figura 51*. En primer lugar se presenta el modelo a ajustar nombrado anteriormente ($y=a+bx$),

junto con las variable elegidas para ajustar el modelo. Después aparece una tabla con los valores estimados para los dos parámetros del mismo: *a* (*Intercept*) y *b* (*Slope*). Éstos son respectivamente 1.27581 y 0.0327699. Esto significa que, si el modelo es aceptable, estimaremos el promedio y los valores individuales de la potencia a partir del peso según la expresión:

HORSEPOWER=1.27581+0.0327699*WEIGHT

Esto es, consideraremos que para la subpoblación de coches que pesan 2300 lbs. la potencia promedio calculada como 1.27581+0.0327699*2300=76.65 caballos es una mejor estimación que la que proporciona la media muestral total. Además, dicho valor será utilizado como predicción individual de cualquier automóvil de 2300 lbs. de peso.

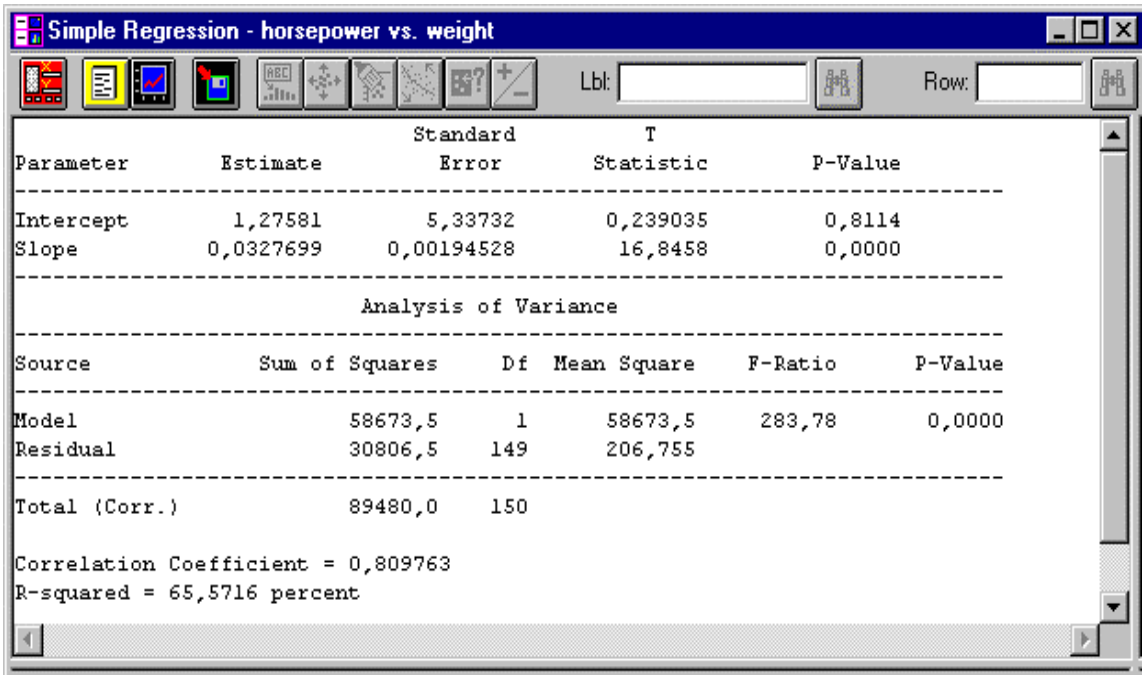


Figura 51

Vemos en la tabla que aparece el error estándar asociado a la estimación del parámetro y otra columna llamada *T-Statistic*. Ésta corresponde al valor de un estadístico calculado para contrastar la hipótesis nula de que el auténtico valor del parámetro poblacional correspondiente sea cero. Vemos que no podemos rechazar la hipótesis para el término independiente (lo que significa que podríamos haber ajustado el modelo $y=bx$) y sí para la pendiente. Es éste último contraste, el de la pendiente, un factor clave en la interpretación de los resultados, como veremos a continuación.

La tabla del Análisis de la Varianza que aparece a continuación ya nos resulta familiar. Descomposición de la variabilidad en la parte que explica el modelo y la residual, y realización de un contraste basado en la distribución F. En el capítulo anterior la hipótesis nula era la igualdad de medias entre grupos, o falta de influencia del factor en la variable dependiente. Esto significaba que el conocimiento de la variable independiente no nos aportaba conocimiento alguno sobre la dependiente. En nuestro caso actual esto se traduce en que la recta de regresión sea constante, que nuestra mejor predicción para la variable dependiente sea la media muestral con independencia del conocimiento de la variable independiente. En definitiva, que la auténtica pendiente poblacional (que estimamos con *b*) sea nula. Por lo tanto, este contraste es equivalente al de la *t* que vimos en el párrafo anterior.

Si el *p-value* obtenido es mayor que, digamos, 0.05, no podremos rechazar la hipótesis nula y concluimos que la variable independiente no nos aporta información alguna. Si, como en nuestro ejemplo, el valor obtenido es significativo, tendremos que analizar cuánta información nos aporta la variable independiente y cuantificar el nivel de certidumbre que podemos tener en las estimaciones y predicciones que realicemos a partir de la recta de regresión construida.

Para ello empezamos fijándonos en dos parámetros que aparecen al final de la pantalla (seguimos en la *Figura 51*): el *coeficiente de correlación* (*correlation coefficient*) y el *coeficiente de determinación* (*R-squared*).

El primero es un número en el intervalo [-1,1] que mide la relación lineal entre ambas variables y su forma conjunta de variar. Valores positivos de este coeficiente significan variación simultánea de ambas variables en el mismo sentido (esto es, a valores más altos de una corresponden valores más altos de la otra), mientras que valores negativos significan variación cruzada (a valores más altos de una corresponden valores más bajos de la otra y viceversa). Valores de este coeficiente más altos en valor absoluto (más cercanos a 1 ó -1) significan una relación lineal más fuerte (puntos más ajustados a la recta) hasta el caso extremo en que los puntos estén totalmente alineados (correlación igual a 1 ó -1).

Podemos calcular coeficientes de correlación para múltiples pares de variables de manera simultánea usando la opción CORRELATION del menú de *Tabular Options* asociado al procedimiento DESCRIBE… NUMERIC DATA… MULTIPLE VARIABLE ANALYSIS.

El coeficiente de determinación es el cuadrado del de correlación, y representa el porcentaje de la variación de la variable dependiente que es explicado por la independiente. Cuando la correlación es 1 ó -1, este coeficiente indica que la variable independiente explica de manera absoluta (100%) el comportamiento de la dependiente.

Estimación y predicción

Ya hemos dicho que utilizaremos la recta de regresión para, dado un valor de la variable independiente, estimar la media de la variable dependiente para la subpoblación que comparte dicho valor. También se usará para predecir valores individuales. En ambos casos el valor será el mismo, aunque lógicamente la predicción individual tendrá mayor error estándar.

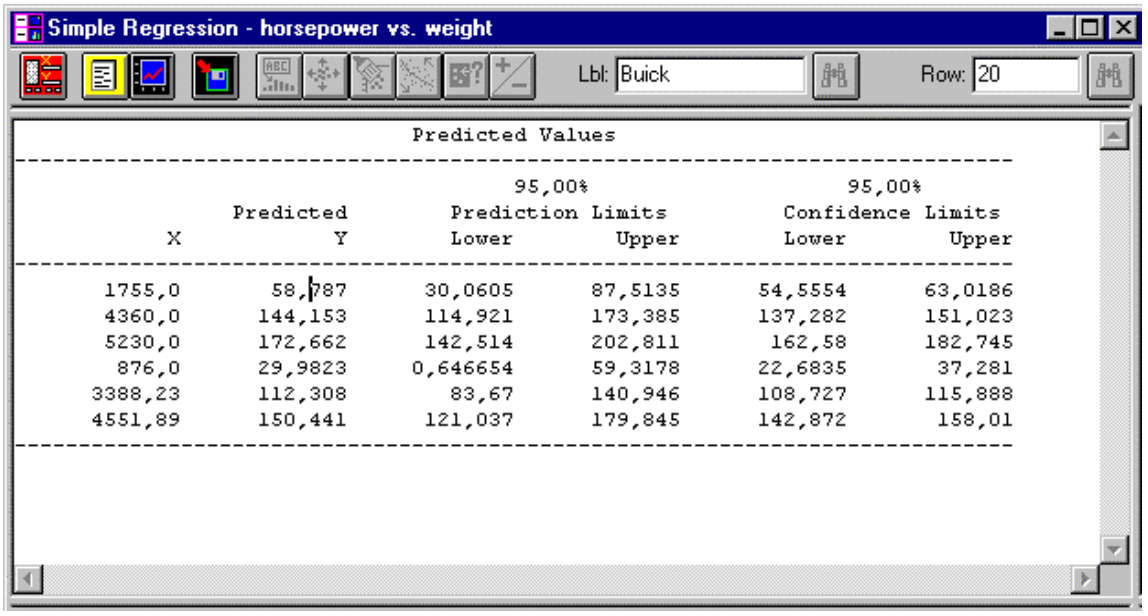


Figura 52

Si en *Tabular Options* seleccionamos FORECASTS, *Pane Options* presenta una pantalla donde escribiremos hasta diez valores para la variable independiente en el campo *Forecast at X:*. Al ejecutar, como podemos ver en la *Figura 52* se nos presentan los valores correspondientes para la variable dependiente calculados a partir de la recta de regresión. Además, dos intervalos de confianza para cada valor, uno considerándolo como estimación de media condicionada (*95% Confidence Limits*) y otro, más amplio, como predicción individual (*95% Prediction Limits*). El nivel de confianza de estos intervalos también puede modificarse en la pantalla de *Pane Options*.

Podremos también representar la nube de puntos muestrales junto con la recta estimada y las franjas de confianza eligiendo la opción PLOT OF FITTED MODEL del menú de *Graphics Options*. La *Figura 53* nos muestra el gráfico obtenido en nuestro ejemplo. Observamos que los bordes de las franjas de confianza no son paralelos a la recta estimada. Esto significa que los intervalos de confianza no tienen siempre la misma amplitud, puesto que las estimaciones hechas a partir de la recta serán mas fiables para valores de la variable independiente cercanos a la media.

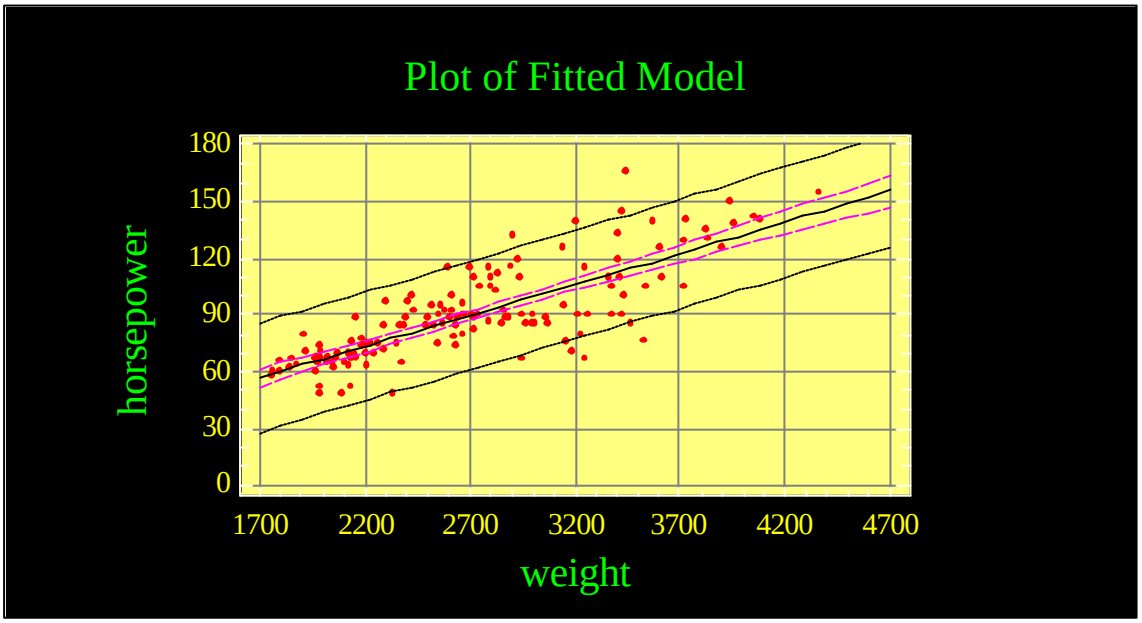


Figura 53

A medida que nos alejamos de ella tenemos menos información disponible en el entorno del valor para el que queremos predecir, y la amplitud de los intervalos de confianza debe, por lo tanto, crecer. A este respecto es importante reseñar que sólo se debe predecir con la recta estimada para valores de la variable independiente dentro del rango de los valores muestrales pues, aunque el modelo se ajuste bien a nuestros datos, no tenemos evidencia de que lo haga fuera de dicho rango.

Hipótesis y validación

El modelo que hemos presentado requiere que se cumplan ciertas hipótesis, la principal de las cuales es la hipótesis de linealidad. Esto es, aplicar el modelo correctamente supone que la naturaleza de la relación entre las variables sea realmente de tipo lineal, al menos en el rango de datos que estemos tratando.

Además, se requiere que la varianza de la variable dependiente se mantenga constante para todos los valores de la independiente y que estas distribuciones condicionadas sean normales.

Para validar las hipótesis es útil el gráfico de puntos de las variables dependiente e independiente, así como el de la variable independiente frente a los residuos. Este último se accede desde la opción PLOT RESIDUALS . En la *Figura 54* vemos el gráfico resultante para nuestro ejemplo. La solución a la violación de alguna de las hipótesis es la transformación de los datos.

En *Analysis Options* tendremos siempre disponibles modelos alternativos al lineal basados en transformaciones de alguna de las dos variables involucradas. La opción COMPARISON OF ALTERNATIVE MODELS nos muestra los coeficientes de correlación y determinación asociados a cada uno de los modelos.

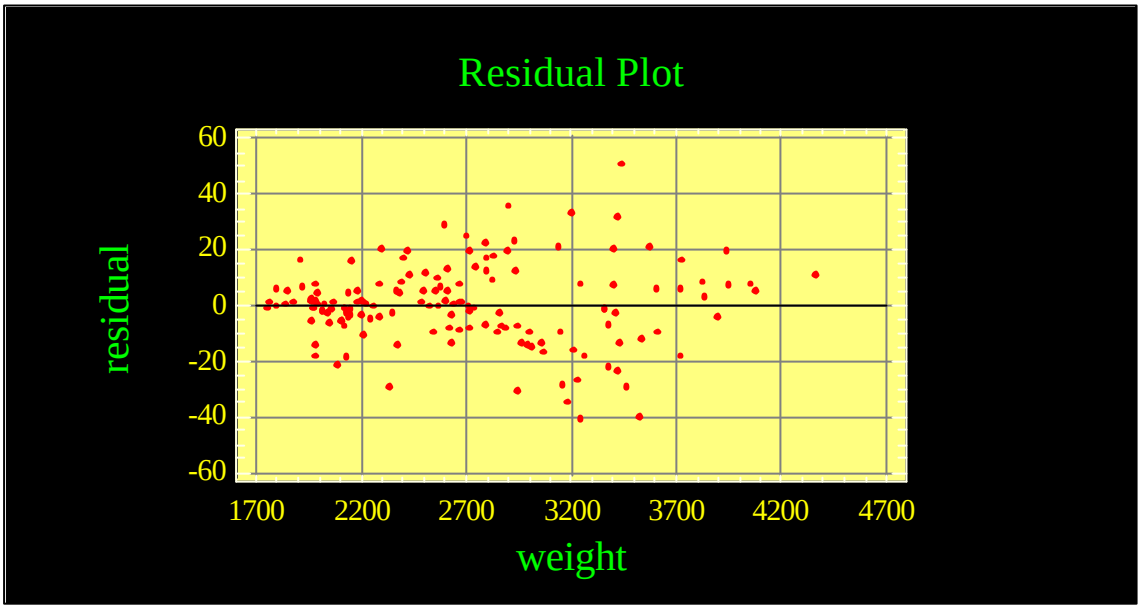


Figura 54

Es muy importante remarcar la influencia que puede tener en el modelo unas pocas (o incluso sólo una) observaciones atípicas. La opción INFLUENTIAL POINTS de *Tabular Options* nos muestra las más influyentes de nuestra muestra. Podemos plantearnos reconstruir el modelo sin alguna(s) de ellas y decidir finalmente cuál es el más adecuado.