Escuela Superior de Ingenieros de San Sebastián
Universidad de Navarra
tecnun

GUÍA DE MINITAB 14

PRÓLOGO

Esta pequeña guía, se ha desarrollado con el objeto de completar la información aportada sobre métodos estadísticos en las aulas a lo largo de las carreras de ingeniería de la Universidad de Navarra.

El enfoque que se ha pretendido dar es, por tanto, sobre el manejo del software para lograr extraer información de un conjunto de datos, sin tratar de explicar cuál es el fundamento de cada una de las técnicas ni el porqué de su interpretación (todo ello, se explica en las clases).

En este sentido, solamente se han desarrollado las técnicas y/o herramientas que se estudian en las asignaturas de Métodos Estadísticos de la Ingeniería, Estadística Industrial y Métodos de Calidad, con el objeto que sirva de ayuda en las prácticas que se realizan durante el transcurso de las mismas.

Quiero señalar que el software MINITAB ofrece además diversas opciones de coloreado, cambios de escalas, cambios de notaciones y representaciones,... para cada técnica gráfica disponible y otras muchas técnicas y herramientas estadísticas que pone a disposición de quien quiera de manera fácil, rápida y asequible, analizar datos. Queda entonces como posibilidad para los alumnos más inquietos, el ahondar en el tan controvertido mundo de la estadística a través de la exploración de este software.

No quiero despedirme sin agradecer el esfuerzo realizado por las alumnas Rosalía De la Rivas e Irene Olaizola en el desarrollo de parte de este manual. Espero que su esfuerzo haya contribuido a realizar una guía práctica y aclaratoria para el resto de los alumnos.

Elisabeth Viles, Septiembre 2004

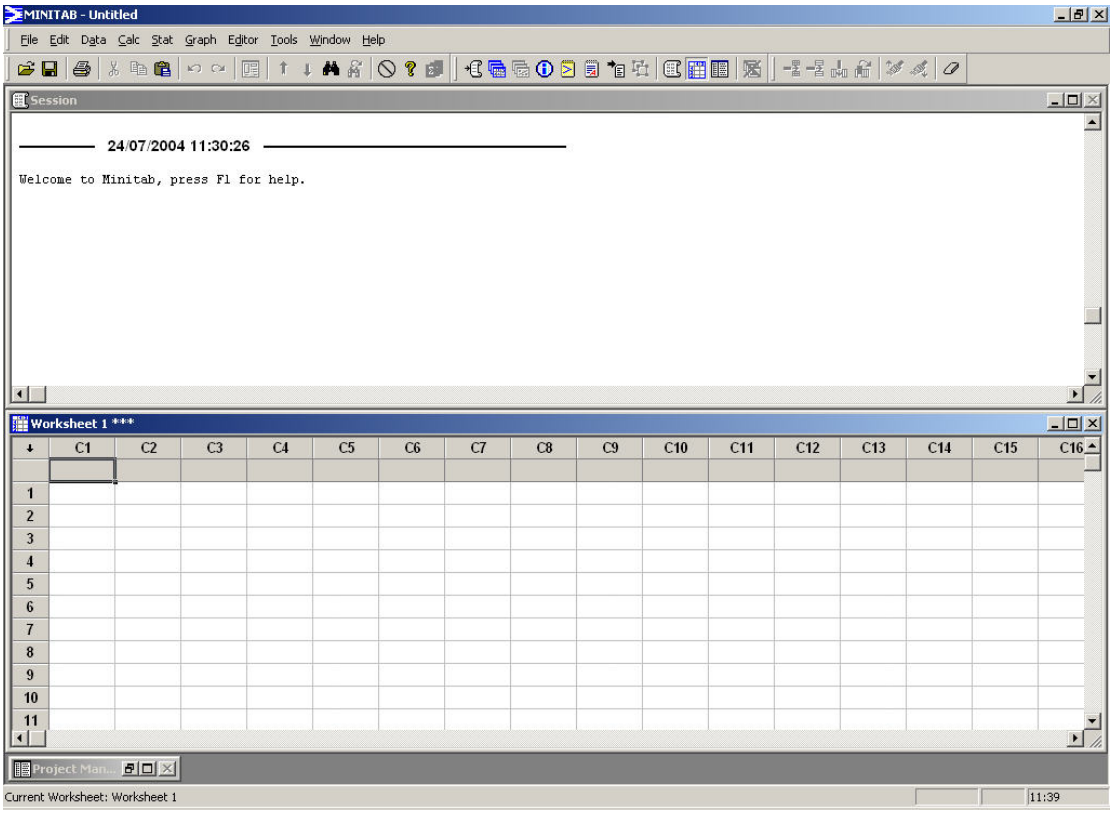
ÍNDICE

1 INTRODUCCIÓN BÁSICA A MINITAB 14

1.1 Ventanas de Minitab

Cuando se abre Minitab aparecen dos ventanas:

1. **Session:** Muestra los resultados de los análisis (tablas, estadísticas…) en formato texto. Además, también permite introducir las órdenes directamente como forma alternativa al uso de los menús (para más información sobre este punto ver la ayuda del programa).

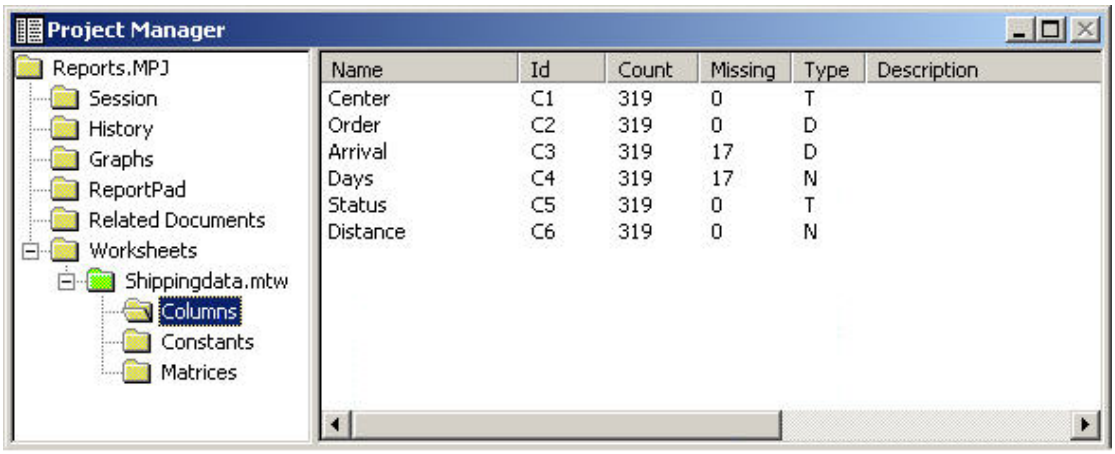


2. **Worksheet:** Estas ventanas contienen filas y columnas (C1, C2..) en las que se introducen los datos del análisis. Se parecen a las hojas de cálculo de Excel. Se puede trabajar con varias hojas de datos a la vez en el mismo proyecto. También se pueden asignar valores a constantes (K1, K2...) y matrices (M1, M2...), pero éstas no aparecen en la hoja de datos.

Otras ventanas del programa:

3. **Graph:** Ventanas de imagen. Se pueden tener hasta más de 200 abiertas simultáneamente.

4. **Project Manager:** Contiene carpetas que permiten navegar por las distintas partes del proyecto. Es muy útil para acceder con rapidez a ventanas que han quedado ocultas.



5. **Report Pad:** Es una herramienta para colocar los datos que nos interesen a modo de informe. Se puede exportar a cualquier editor de texto o imprimirlo directamente. Para incluir algún gráfico o texto de Session en el Report Pad basta con pulsar con el botón derecho del ratón sobre el objeto en cuestión y seleccionar *Append to Report*

1.2 Introducir datos

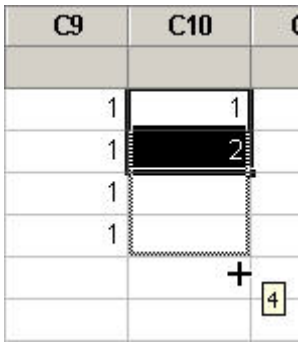
Las variables pueden definirse como texto, numéricas o de tiempo. Minitab las define automáticamente con los primeros valores que se introducen la hoja de datos. Hay varias maneras de introducir datos en Minitab.

	C1-T	C2-D	C3-D	C4	C5-T	C6	C7	C8
	Central	Orden	LLegada	Días	Estado	Distancia		
1	Centro	3/3/2003 8:34	3/7/2003 15:21	4,28264	En hora	255		
2	Este	3/3/2003 8:35	3/6/2003 17:05	3,35417	En hora	196		
3	Oeste	3/3/2003 8:38	*	*	Retraso	299		
4	Centro	3/3/2003 8:40	3/7/2003 15:52	4,30000	En hora	205		
5	Este	3/3/2003 8:42	3/9/2003 14:48	6,25417	Tarde	250		
6	Oeste	3/3/2003 8:43	3/8/2003 15:45	5,29306	En hora	93		
7	Centro	3/3/2003 8:50	3/7/2003 10:02	4,05000	En hora	189		
8								

Se pueden escribir directamente en la hoja de datos. Normalmente cada columna es una variable y cada fila un valor particular de la variable. Se puede escribir el nombre de la variable debajo de C1, C2... Si la flecha del extremo superior izquierdo

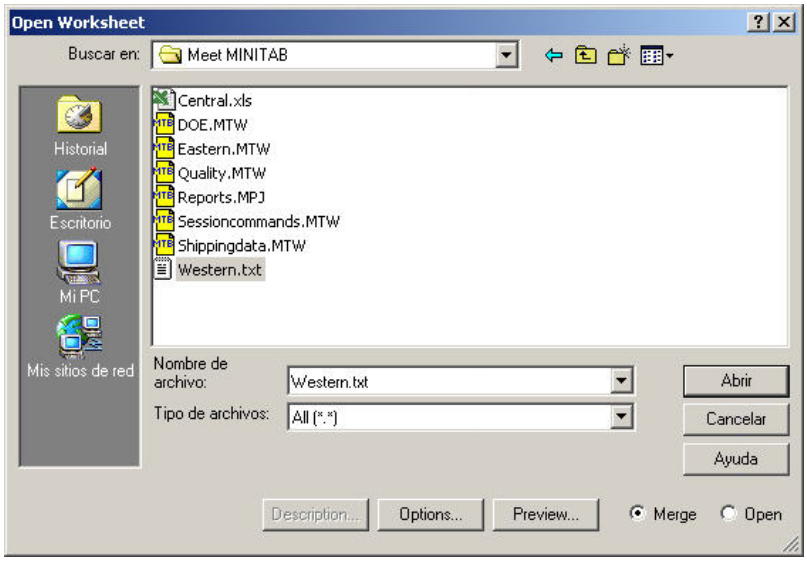
señala hacia abajo, cuando se pulsa <enter> se selecciona la celda inmediatamente inferior; y cuando se pulsa <control> <enter> se selecciona la primera celda de la columna siguiente. Se puede cambiar la dirección de la flecha pulsándola.

Una opción muy útil es utilizar el arrastre (*autofill*) como en Excel. Si, por ejemplo, se escribe un 1 en una celda y se arrastra la cruz hacia las celdas inferiores, éstas se rellenarán también con 1. Si escribiésemos una sucesión sencilla. 1,2,3... y arrastrásemos, la sucesión continuaría en las siguientes celdas.



El Minitab reconoce archivos de datos creados por otros sistemas o aplicaciones. Así, pueden importarse datos desde Excel o Notepad.

Merge: Sitúa las nuevas columnas a continuación de las que ya existían.



Open: Abre el documento en una nueva hoja de datos.

Preview: Aparecen las cien primeras entradas de la hoja de datos. Si no se desea incluir alguna de las variables basta con eliminar el nombre de su columna C1, C2... Permite también cambiar el tipo de variable (texto, numérico, tiempo). Si no reconoce los caracteres los marcará como entrada desconocida y se representa con un asterisco (*).

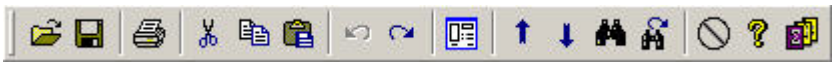
	C1	C2	C3
	Text	Numeric	Numeric
1	Life	6	19
2	Grape Nuts	3	23
3	Super Sugar Crisp	2	26

Options: Se puede ordenar a Minitab que interprete la primera línea de datos, por ejemplo, como la línea que contiene los nombres de las variables.

Nota: Minitab interpreta valores vacíos o celdas de data-time como valores perdidos, y se representan por asterisco (*). Sin embargo, cuando la última fila de una columna de datos de un archivo de texto aparece vacía, Minitab la deja vacía en lugar de ponerle el asterisco.

1.3 Barras de herramientas

Por defecto encontraremos tres barras de herramientas:

Standard: 

La novedad de esta barra de herramientas frente a la de otros programas es el botón "Edit last command"  . Con él podemos modificar las propiedades de la última orden que se le ha dado al programa.

Project Manager: 

Es un acceso directo a las distintas vistas predefinidas por el programa:

Session  , Worksheet  , Project Manager  , y combinaciones de Project Manager con Session  , Worksheet  , y gráficos  . Además podemos acceder a una breve información a cerca de las variables utilizadas  y de las últimas órdenes dadas  .

Worksheet: 

Son accesos directos al menú Editor. Insertar celda, fila, columna, mover columnas, ir a la columna pegada anterior, ir a la columna pegada siguiente, borrar celdas.

1.4 Guardar datos

Desde **File** se accede a las opciones como crear un nuevo fichero, abrir uno existente, grabar, leer datos de otras aplicaciones, imprimir…

Hay varias opciones para grabar. Se pueden guardar las hojas de datos o Worksheets (extensión *.MTW), los gráficos (varias extensiones, entre ellas la propia del Minitab: *.MGF), o el proyecto entero (*.MPJ). Un archivo sólo se puede

recuperar de la forma como ha sido grabado. Si se ha grabado como Worksheet se recupera como Worksheet. Sucede lo mismo para los Project.

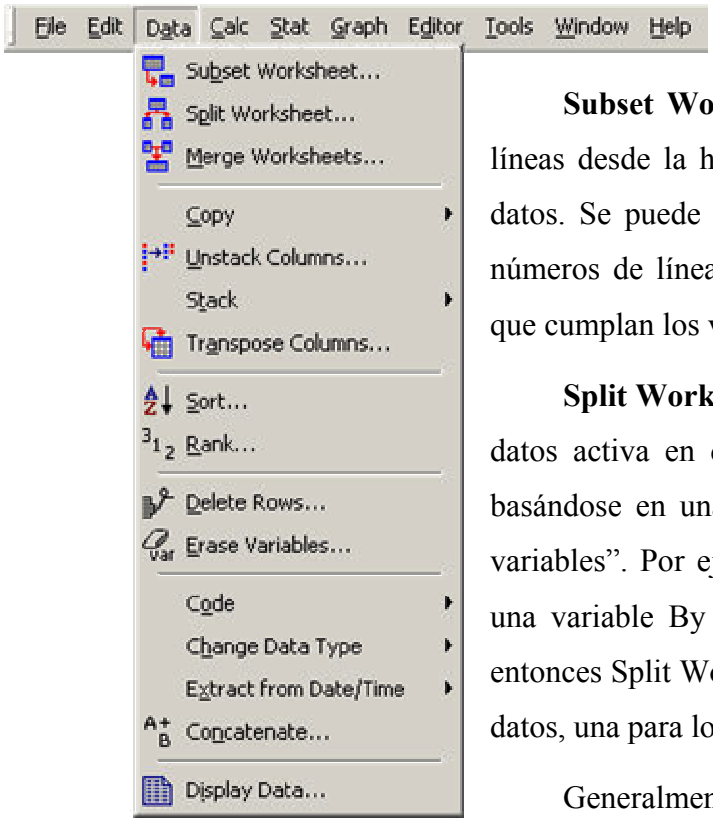
1.5 Edición de datos

Los datos se introducen en el la hoja de datos y se editan de manera similar a la hoja de cálculo de Excel. Con las opciones **Editor/Insert Columns** y **Editor/Insert Rows** se pueden añadir nuevas variables y nuevos casos al archivo. El programa sitúa la nueva variable o el caso en la posición inmediatamente anterior a la celda activa. Puede modificarse cualquier dato introducido, así como elegir las opciones del menú **Edit:** cortar, copiar, pegar, eliminar…

El menú **Editor/Column** permite hacer otros cambios como modificar el ancho de columna o esconder algunas variables. A menudo es más rápido editar las celdas con el menú que surge al pulsar el botón derecho del ratón sobre ellas.

2 Manipular datos

2.1 Menu Data



Subset Worksheet se usa para copiar ciertas líneas desde la hoja de datos activa a otra hoja de datos. Se puede hacer el Subset en función de los números de línea o de alguna condición específica que cumplan los valores.

Split Worksheet se usa para separar la hoja de datos activa en dos o más nuevas hojas de datos basándose en una o más variables del campo "By variables". Por ejemplo, si se ejecuta basándose en una variable By que contiene los valores sí y no, entonces Split Worksheet creará dos nuevas hojas de datos, una para los sí, y otra para los no.

Generalmente Split Worksheet sólo es útil cuando crea pocas nuevas hojas de datos. Sin embargo no hay límite del número de nuevas hojas que se pueden crear.

Merge Worksheet se usa para combinar dos hojas de datos abiertas. Duplica y a continuación puede combinar la información de las dos hojas de datos. La opción por defecto combina las hojas de datos una junto a la otra y con los atributos que ya tenían. Se puede usar "By Columns" para combinar la hoja de datos de acuerdo con el orden y longitud de una o más columnas. Además se puede especificar si se quiere o no incluir valores que no se corresponden (unmatched observations), valores que faltan (missing observations, para incluirlos hay que pulsar en "incluye missing as by level") o valores repetidos (multiple observations) de cualquiera de las hojas originales. También se pueden especificar las columnas que se van a incluir y las que no.

Unstack Columns sirve para separar bloques de columnas en otras más cortas. Por ejemplo, en la siguiente figura vemos cómo se desglosan las ventas de Denver Boston y Seattle en otras columnas con Unstack. **Stack** es el proceso inverso.

Sales	Store	Sales_Boston	Sales_Denver	Sales_Seattle
52	Denver	36	52	63
36	Boston	32	46	71
63	Seattle	35	51	68
46	Denver	29	50	66
32	Boston			
71	Seattle			
51	Denver			
35	Boston			
68	Seattle			
50	Denver			
29	Boston			
66	Seattle			

Traspose columns permite trasponer un grupo de filas y columnas para reordenar los datos. Para la imagen de ejemplo se ha hecho: en **Transpose the following columns**, se escribe *Lyn, Bill, Sam,* y *Marie*; en **Store transpose**, se elige **After last column in use**; y en **Create variable names using column**, se introduce *Task*.

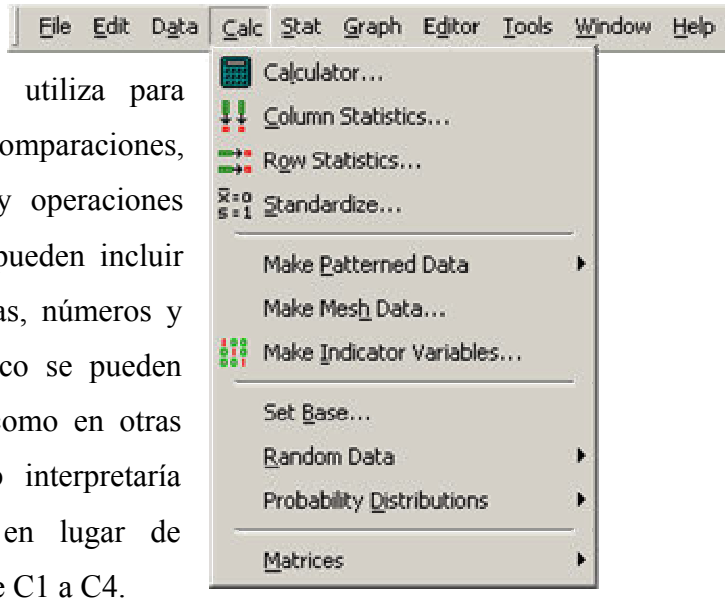
C1-T	C2	C3	C4	C5	C6-T	C7	C8	C9
Task	Lyn	Bill	Sam	Marie	Labels	Pushups	Pullups	Situps
Pushups	50	69	70	57	Lyn	50	66	73
Pullups	66	85	81	76	Bill	69	85	88
Situps	73	88	95	79	Sam	70	81	95
					Marie	57	76	79

Sort ordena alfabética o numéricamente las variables que se le indiquen. Y **Rank**, en cambio, indica en qué posición estarían en caso de que se ordenasen con **Sort.**

A veces será conveniente recodificar las variables. No consiste exactamente en cambiar el tipo de formato a las celdas, sino en reinterpretarlas con el mismo u otro formato. Desde **Data/Code** se pueden resolver ejemplos como los siguientes: cambiar el nombre de los meses (enero, febrero…) por números en las fechas (01/01/04, 01/02/04…) o hacer que todas las puntuaciones entre 85 y 100 reciban un sobresaliente, las de 65 a 85 un notable y las de 50 a 65 un aprobado.

Las variables pueden definirse como texto (T), numéricas (N) o de tiempo (D). Minitab las define automáticamente con los primeros valores que se introducen la hoja de datos. De todas formas, siempre cabe la posibilidad de cambiarlas de un tipo a otro con el menú **Data/Change Data Type**. Para poder realizar los cambios deben ser coherentes: por ejemplo, si se intenta cambiar el nombre de una calle a numérico aparecerá un asterisco (*) en la conversión indicando que el valor falta.

2.2 Menu Calc



Calculator: Calculator se utiliza para hacer operaciones aritméticas, comparaciones, operaciones lógicas, funciones y operaciones por columnas. Las expresiones pueden incluir columnas, constantes almacenadas, números y texto, pero no matrices. Tampoco se pueden utilizar expresiones abreviadas como en otras ventanas. Por ejemplo, Minitab interpretaría C1-C4 como C1 menos C4 en lugar de interpretarlo como las variables de C1 a C4.

Make Patterned Data: Se puede usar esta opción para crear una gran variedad de patrones de datos. No es tan rápido como el *Autofill* pero permite hacer con más facilidad listas largas de datos con valores repetidos. Sus opciones son:

Simple Set of Numbers: llenar una columna con un patrón de números igualmente espaciados.

Arbitrary Set of Numbers: llenar una columna con un patrón de números que no están igualmente espaciados.

Text Values: llenar una columna con un patrón de valores de texto

Simple Set of Date/Time Values: llenar una columna con una serie de fechas o datos temporales.

Arbitrary Set of Date/Time Values: llenar una columna con un patrón de fechas o datos temporales.

Make Indicator Variables: En la opción *Indicator variables for* se introduce la columna que contiene los valores fijos. Por ejemplo una columna que contenga las palabras "Blue", "Green", "Brown" y "Hazle". En la columna *Store results in* se introducen tantas columnas como variables halla, sin contar los valores perdidos (*). En el ejemplo serían cuatro columnas, una por cada color. **Make Indicator Variables** creará, en este caso, cuatro columnas con ceros y unos. Si en una fila está escrita la palabra "Blue", en la columna correspondiente a "Blue" que se crea pondrá un uno y en el resto ceros; si está escrito "Green" en la columna correspondiente a ese color pondrá un uno y en el resto ceros, y así sucesivamente.

↓	C1-T	C2	C3	C4	C5
	Eyecolor				
1	Brown	0	1	0	0
2	Blue	1	0	0	0
3	Hazel	0	0	0	1
4	Green	0	0	1	0
5		*	*	*	*

Set Base permite modificar la semilla de generación de números aleatorios.

Random Data puede usarse para crear filas o columnas de números aleatorios con hasta 24 distribuciones de datos diferentes.

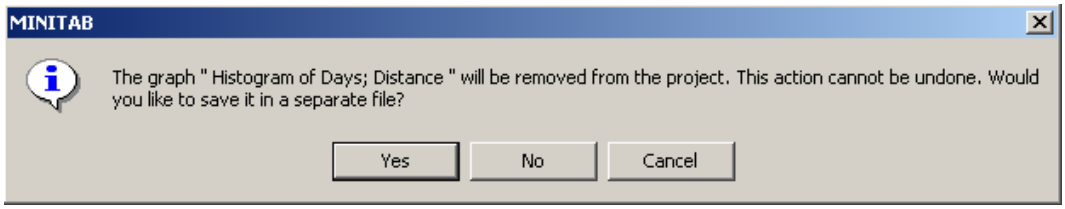
Probability Distributions permite calcular la función de masa (casos discretos), valores de la función de densidad (casos continuos) y funciones de distribución de las distribuciones de probabilidad más usuales.

2.3 Gráficos

Los gráficos son una parte importante de los proyectos y suele resultar conveniente utilizarlos en cualquier estudio estadístico porque permiten intuir algunos resultados. Es importante hacer notar que cuando guardemos un estudio estadístico realizado en MINITAB como proyecto (.MPJ), se guardan con él todos los gráficos que se han realizado en dicho estudio y están en ese momento abiertos. No ocurre así, si el estudio se guarda como documento de datos o worksheet (.MTW)

Se pueden abrir gráficos guardados en Minitab Graphics Format (.MGF), así como gráficos incluidos dentro de proyectos, a través de la ruta **File/Open Graph**.

Cuando se cierra un gráfico, éste se borra del proyecto; el programa da la posibilidad de guardarlo como un archivo aparte eligiendo la opción 'Yes' en la siguiente ventana.



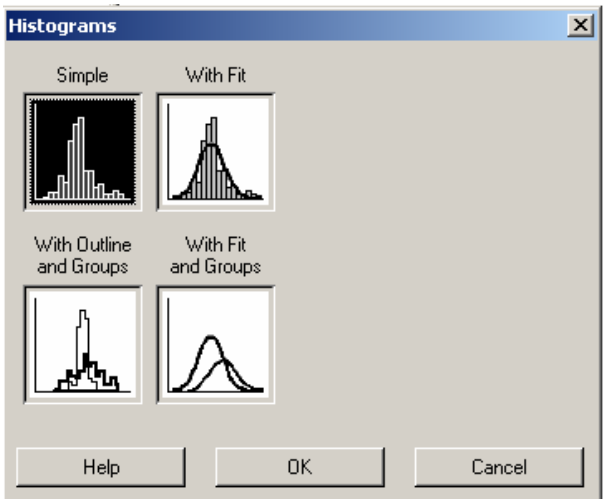
Minitab permite realizar una gran cantidad de gráficos y modificar muchas características tanto antes como después de crearlo. Aquí veremos sólo los aspectos que más nos interesan.

2.3.1 Histogramas

Se utilizan para representar la variabilidad de un conjunto de datos, utilizando como valores representativos del conjunto de datos, su valor medio y su desviación típica.

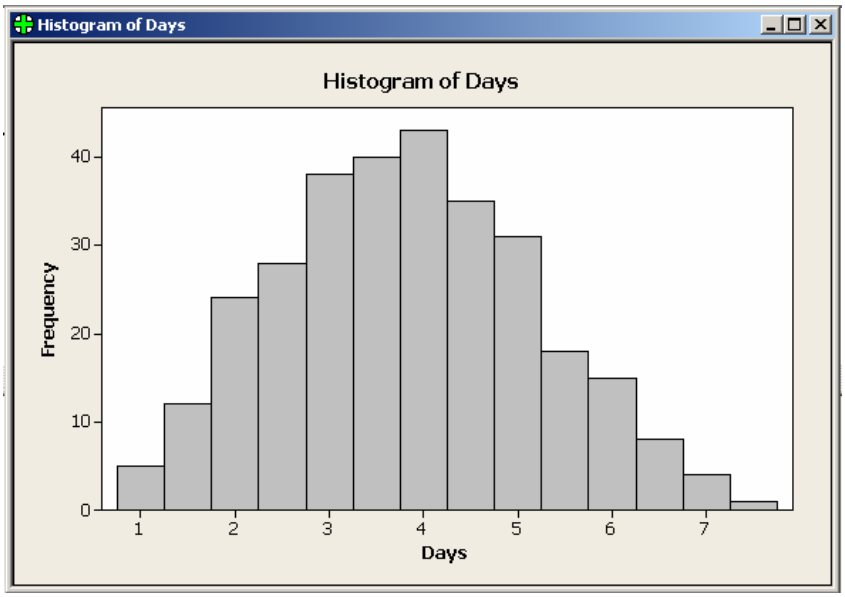
Representan, por tanto, la frecuencia de los datos de una variable dividida en diferentes intervalos. Se logran según la ruta **Graph/Histogram.**

Existen diferentes tipos de histogramas:



Simple:

En la caja de texto de la izquierda se muestran las variables entre las que se puede elegir. Hay que rellenar la caja "Graph Variables"; tras seleccionarla, se clica dos veces sobre las variables de las que se quiera su histograma. Después de personalizarlo con las múltiples opciones existentes, se pulsa "ok". En este caso, para visulizar el histograma de la variable Days, se obtiene:

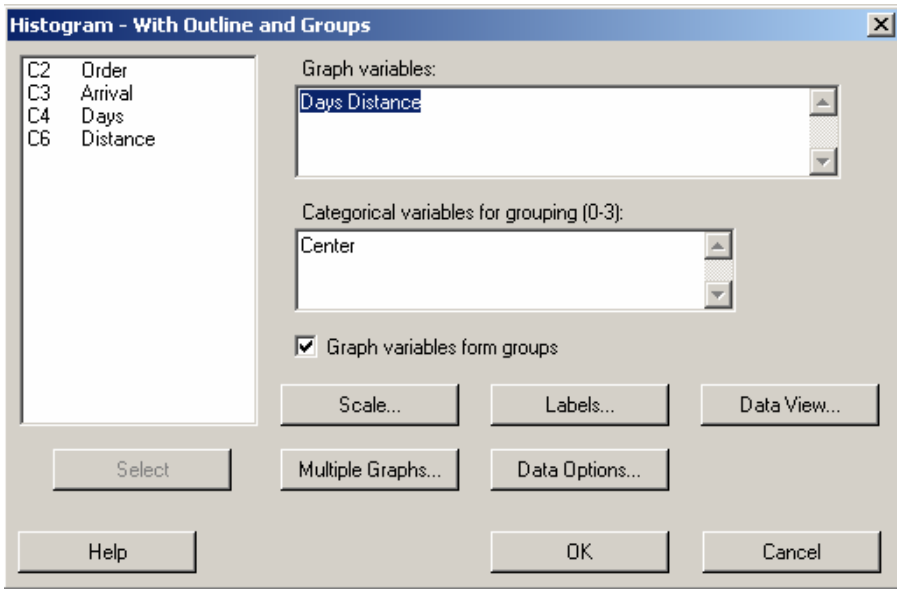


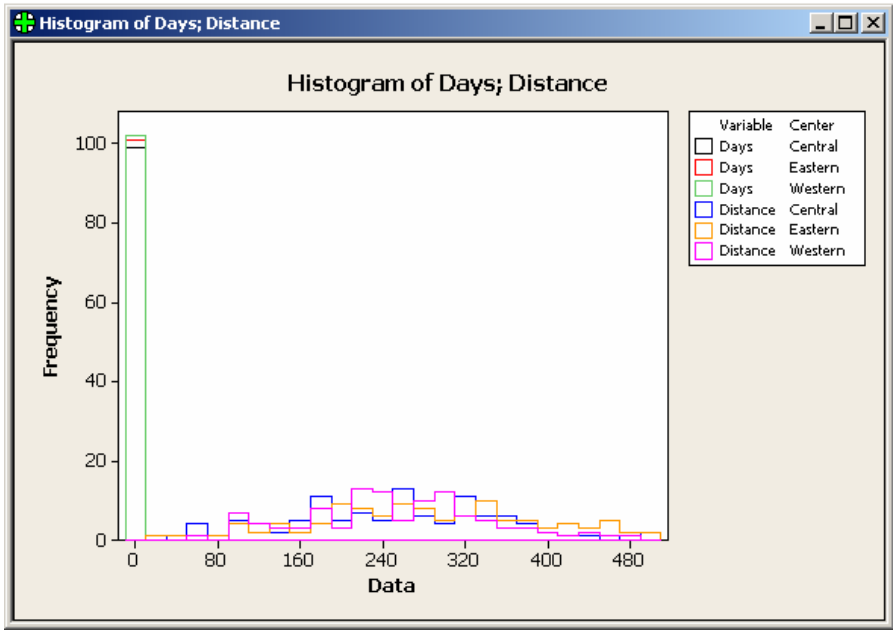
With Fit:

Se sigue el proceso similar al anterior. La única diferencia es que aparecerá la curva normal de distribución sobreimpreso en el histograma.

With Outline and Groups:

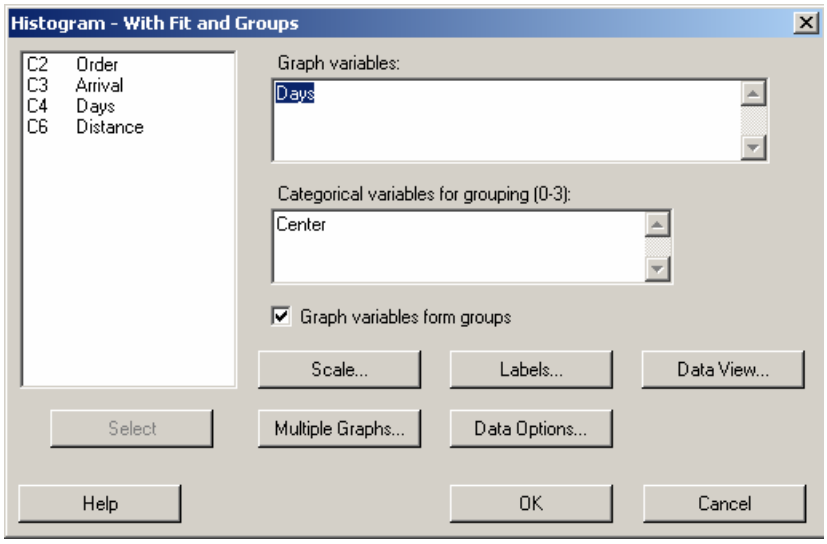
La diferencia con los casos anteriores, es que se puede agrupar las variables utilizadas según alguna otra. Por ejemplo, si vamos a representar el número de días que tarda en realizarse un envío, así como la distancia al punto de envío, se pueden agrupar dichas variables en función del centro desde donde se envíen (en este caso Eastern, Western y Central). El resultado es el siguiente:

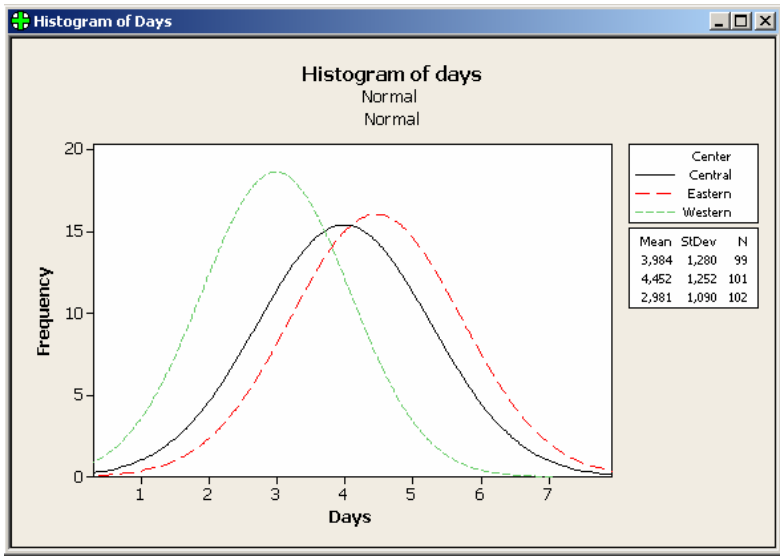




With Fit and Groups:

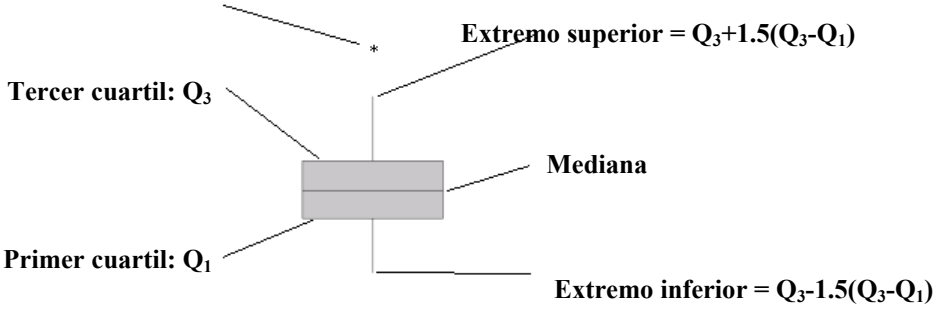
La idea es la misma que en el caso anterior, pero en vez de un diagrama de barras, obtenemos la curva de distribución normal.



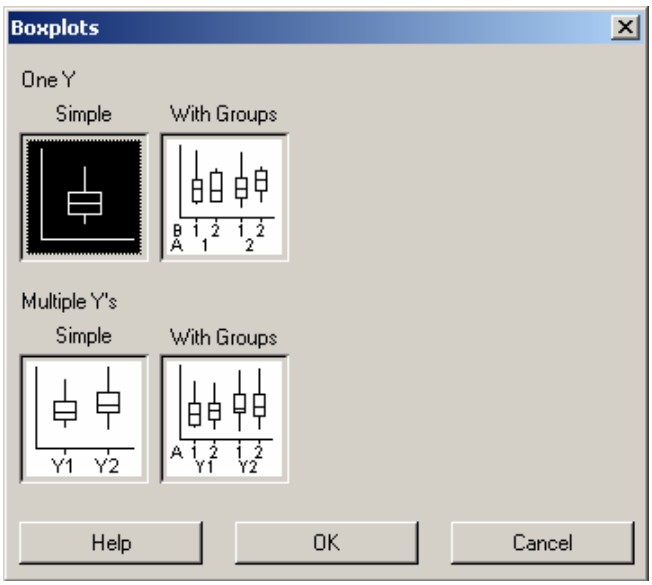


2.3.2 Diagramas de Caja

Un diagrama de caja se utiliza también para visualizar la variabilidad de un conjunto de datos pero esta vez utilizando como valores representativos de estos datos, lo que se llaman valores de posición: menor, mayor y cuartiles. El siguiente dibujo especifica sobre la caja estos valores:

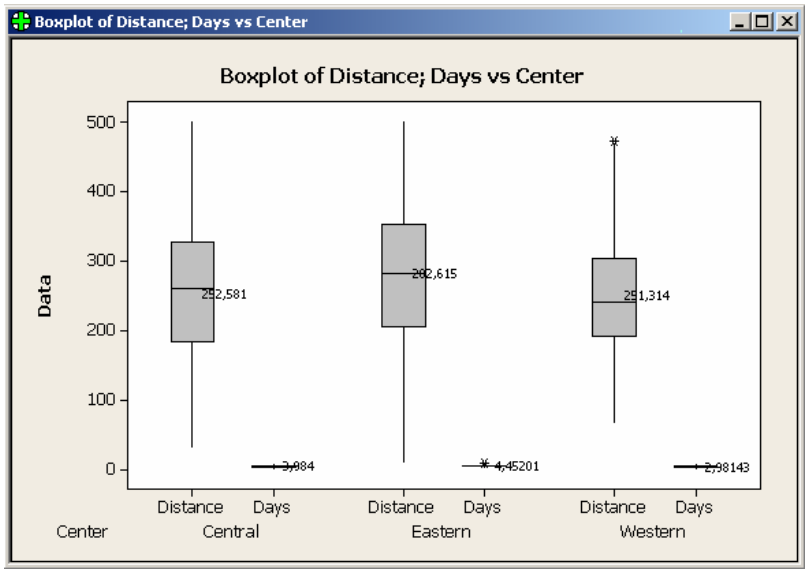


Esta gráfica se logra según la ruta **Graph/Boxplot**



Existen cuatro tipos diferentes que se consiguen de manera análoga. En el ejemplo vemos cómo realizar uno del tipo *Multiple Y's–With Grups.*

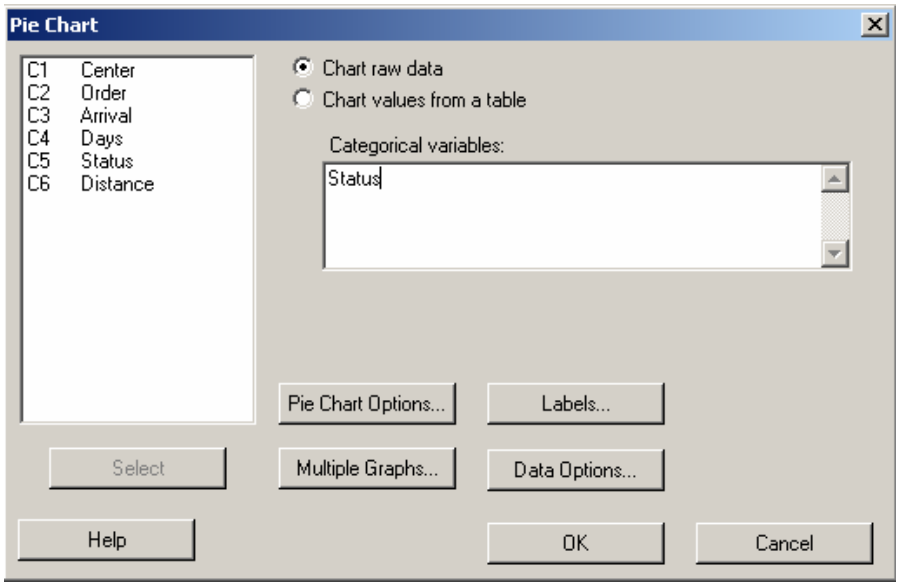
En *Graph Variables* se introducen las variables que se desea representar, por ejemplo Distancedays, y en *Categorical variables for grouping* la variable en función de la cual se agruparán los datos, Center. El resultado es:



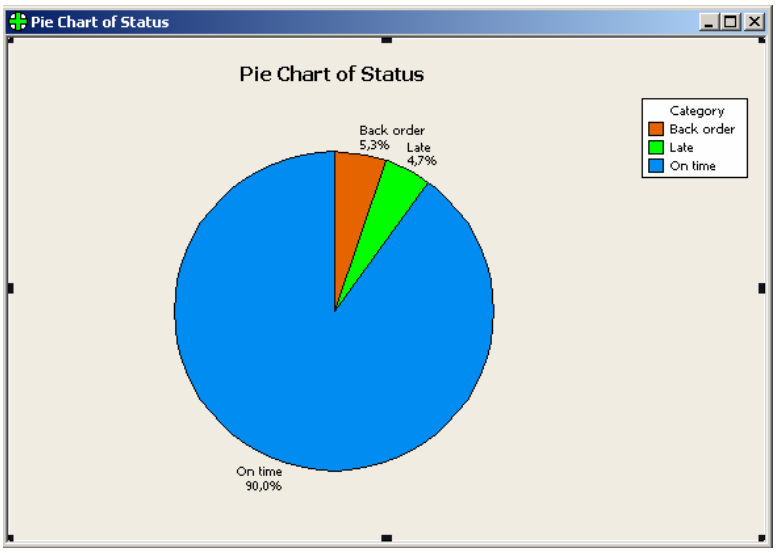
2.3.3 Diagramas Circulares

Se utilizan para mostrar proporciones de valores que toman las variables en comparación con la muestra total. Se logran a través de la ruta **Graph/Pie Chart.**

Si se elige la opción *Chart raw data*:



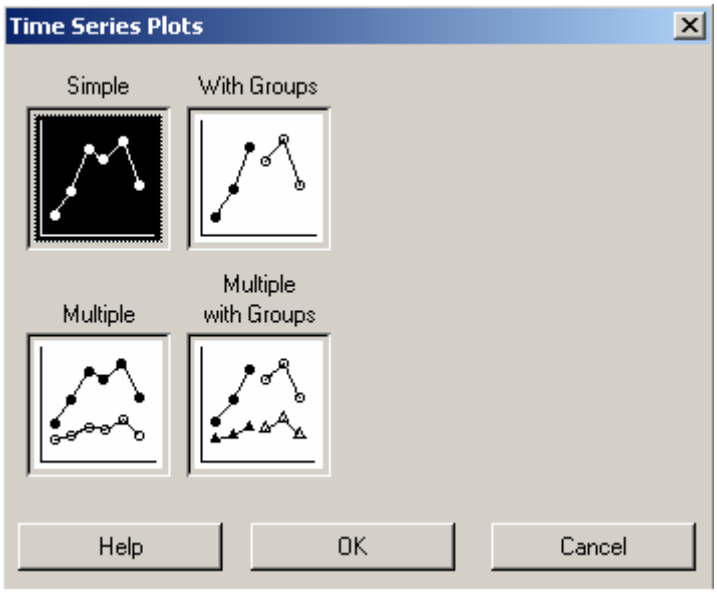
En *Categorical Variables* se introduce la variable que se desea analizar. El resultado es:



2.3.4 Diagramas temporales

Se utilizan cuando los datos están tomados a intervalos iguales de tiempo y están en orden cronológico en la hoja.

Se logran siguiendo la ruta **Graph/Time Series Plot**. Existen cuatro tipos diferentes:



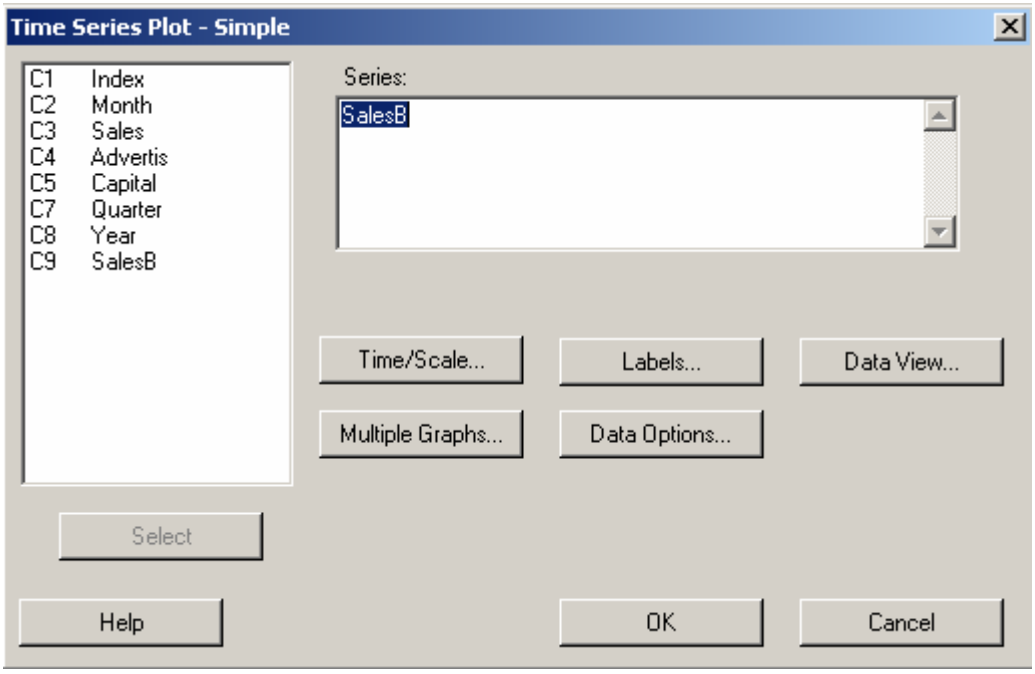
Se realizará uno simple; el resto son análogos.

Se quiere visualizar las ventas realizadas en cada uno de los trimestres de los años 2001 a 2003. La tabla de datos de que se dispone sería la siguiente:

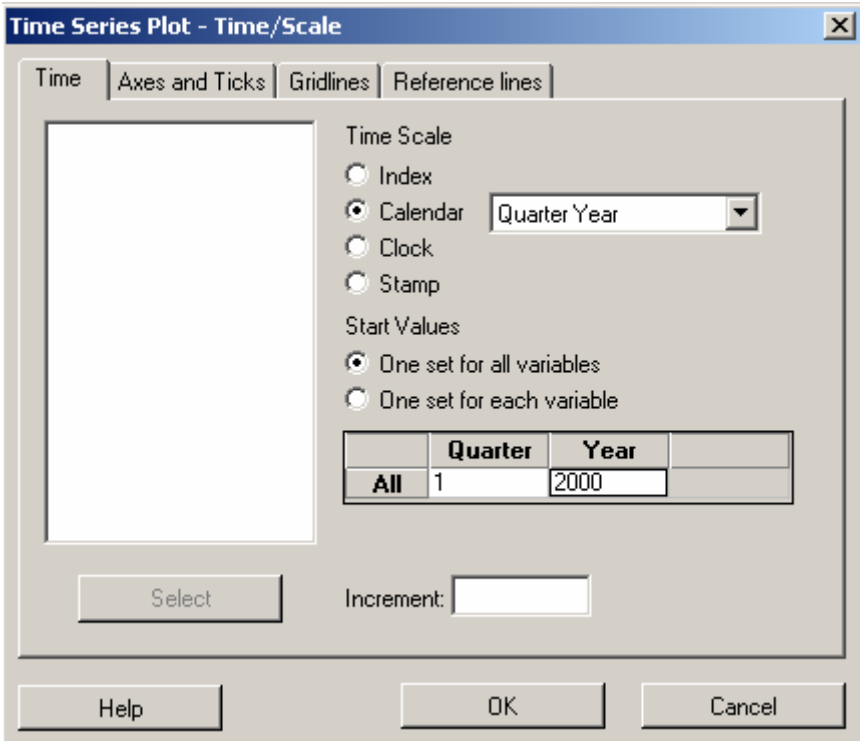
Newmarket.MTW ***

↓	C1	C2-D	C3	C4	C5	C6-T	C7	C8	C9	C10-T
	Index	Month	Sales	Advertis	Capital	AdAgency	Quarter	Year	SalesB	Date
1	1	enero	210	30	14	Omega	1	2000	100	1Q00
2	2	febrero	205	25	18	Omega	2	2000	120	2Q00
3	3	marzo	202	18	20	Alpha	3	2000	180	3Q00
4	4	abril	245	55	24	Alpha	4	2000	183	4Q00
5	5	mayo	237	32	31	Alpha	1	2001	143	1Q01
6	6	junio	290	27	35	Alpha	2	2001	151	2Q01
7	7	julio	299	23	36	Omega	3	2001	199	3Q01
8	8	agosto	345	60	41	Omega	4	2001	211	4Q01
9	9	septiembre	326	34	42	Alpha	1	2002	165	1Q02
10	10	octubre	355	30	46	Alpha	2	2002	193	2Q02
11							3	2002	205	3Q02
12							4	2002	235	4Q02

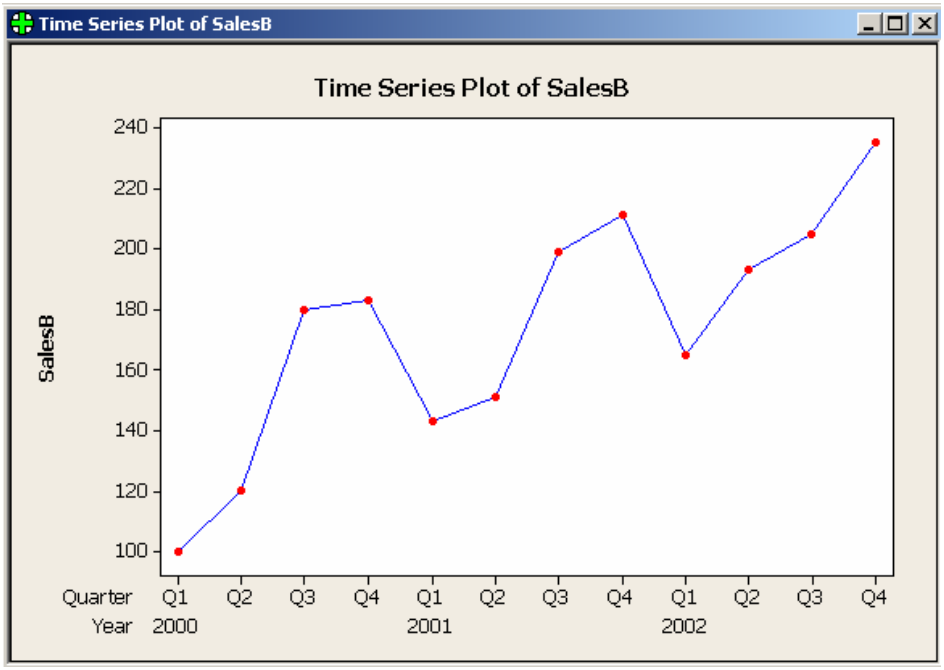
Siguiendo la ruta mencionada anteriormente y tras pulsar en *Simple* se obtiene:



Se selecciona la columna *SalesB* para representarla. A continuación se elige la opción *Time/Scale* para personalizar el tipo de eje temporal que se desea establecer.



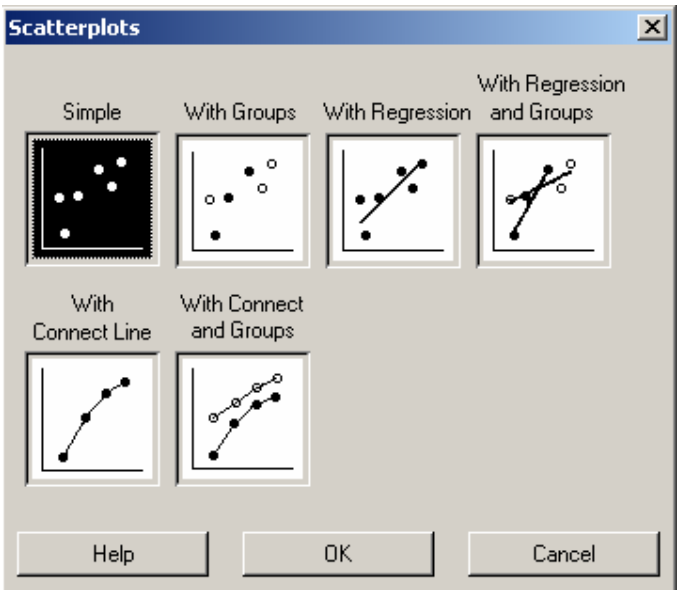
Al elegir *Quarter Year* se dispone que cada dato del número de ventas va a referirse a un trimestre de un año en orden cronológico. En *Start Values* se eligen los valores desde los que se van a inicializar el trimestre y el año; en este caso trimestre primero del año 2000. El resultado es el siguiente:



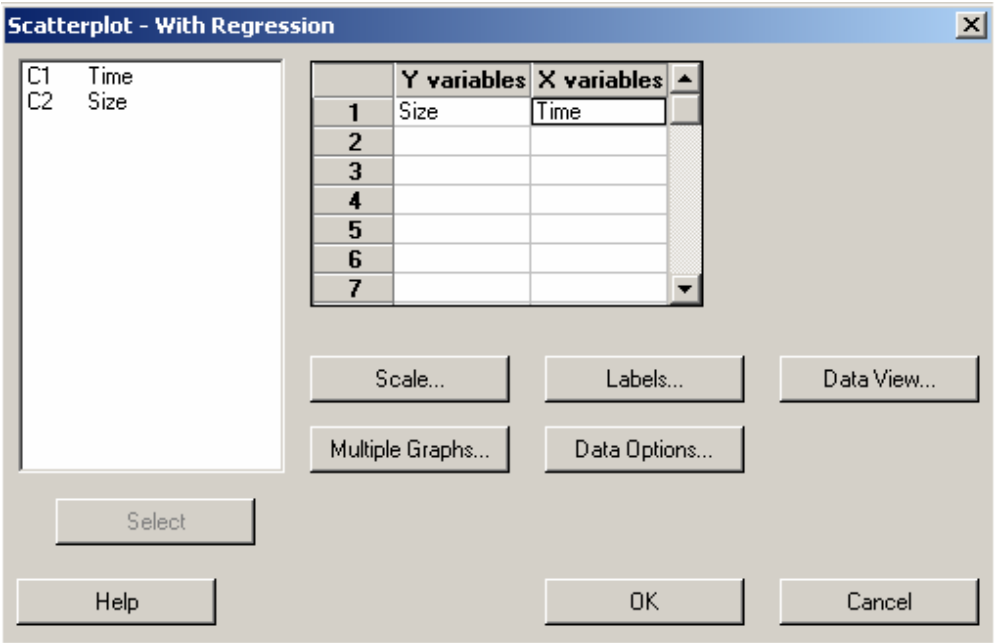
2.3.5 Diagramas de dispersión

Se utilizan para ilustrar la relación existente entre dos variables, representando una frente a la otra. Asimismo, son útiles para representar variables frente al tiempo, cuando los datos no están ordenados cronológicamente.

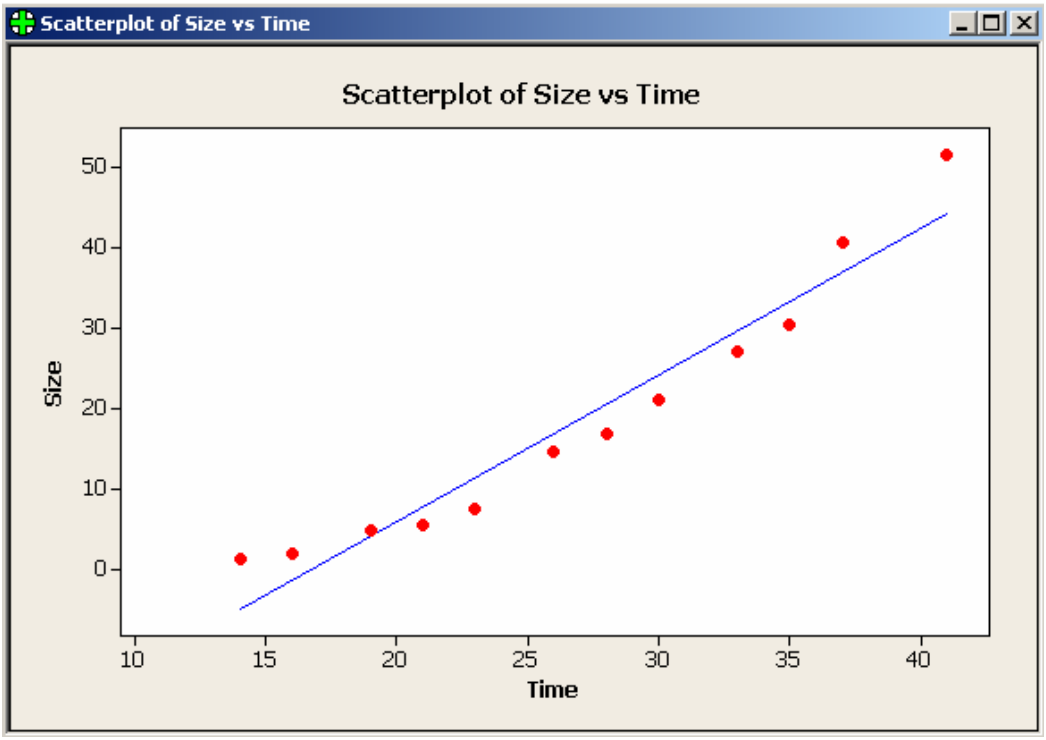
Se logran siguiendo la ruta **Graphs/Scatterplots**. Existen diferentes tipos:



Se realizará un gráfico *With Regresion* en el que se verá la relación existente entre el tiempo de desarrollo de un tumor y el tamaño de éste. Tras rellenar el cuadro de diálogo de la forma mostrada:



Se obtiene:



Véase la importancia de situar correctamente las variables en los ejes X o Y respectivamente. La recta de regresión varía, ya que se ajusta por mínimos cuadrados.

2.3.6 Paretos

Los paretos son un tipo de gráficos en los que el eje horizontal representa categorías de interés frente a la frecuencia (relativa o absoluta) de aparición de estas categorías en los datos. El gráfico ordena estas categorías de acuerdo a sus frecuencias, de mayor a menor, lo que permite al usuario utilizar el pareto para determinar qué categorías de las estudiadas pueden ser vitales y cuáles triviales. La línea de porcentaje acumulativa ayuda a juzgar la contribución de cada categoría.

Veamos un ejemplo realizando un pareto de los tipos de defectos encontrados en una línea de montaje de circuitos electrónicos:

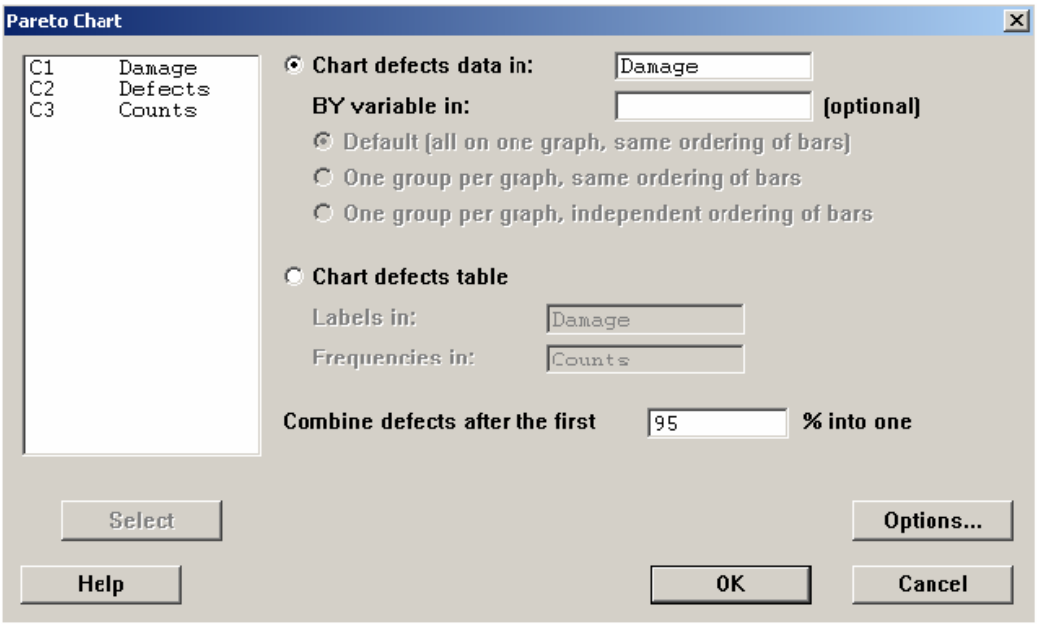
La ruta que hay que seguir es **Stat/QualityTool/ParetoChart**. Sobre la siguiente tabla de defectos:

Exh_qc.MTW ***

	C1-T	C2-T	C3
	Damage	Defects	Counts
1	Scratch	Missing Screws	274
2	Scratch	Missing Clips	59
3	Bend	Defective Housi	19
4	Chip	Leaky Gasket	43
5	Dent	Scrap	4
6	Scratch	Unconnected Wir	8
7	Chip	Missing Studs	6
8	Scratch	Incomplete Part	10

Existen dos tipos posibles maneras de hacer el pareto en función de cómo tengamos los datos:

Chart defects data in: Los nombres de los defectos se han colocado en una columna, y cada defecto aparece tantas veces como ha sido encontrado. En este caso:



El resultado es:

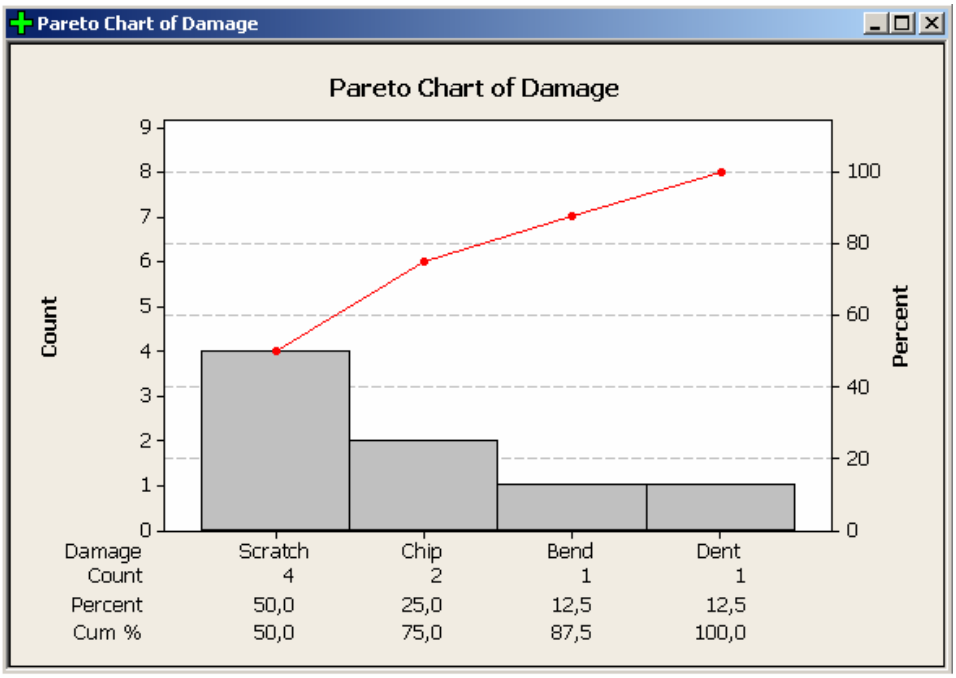
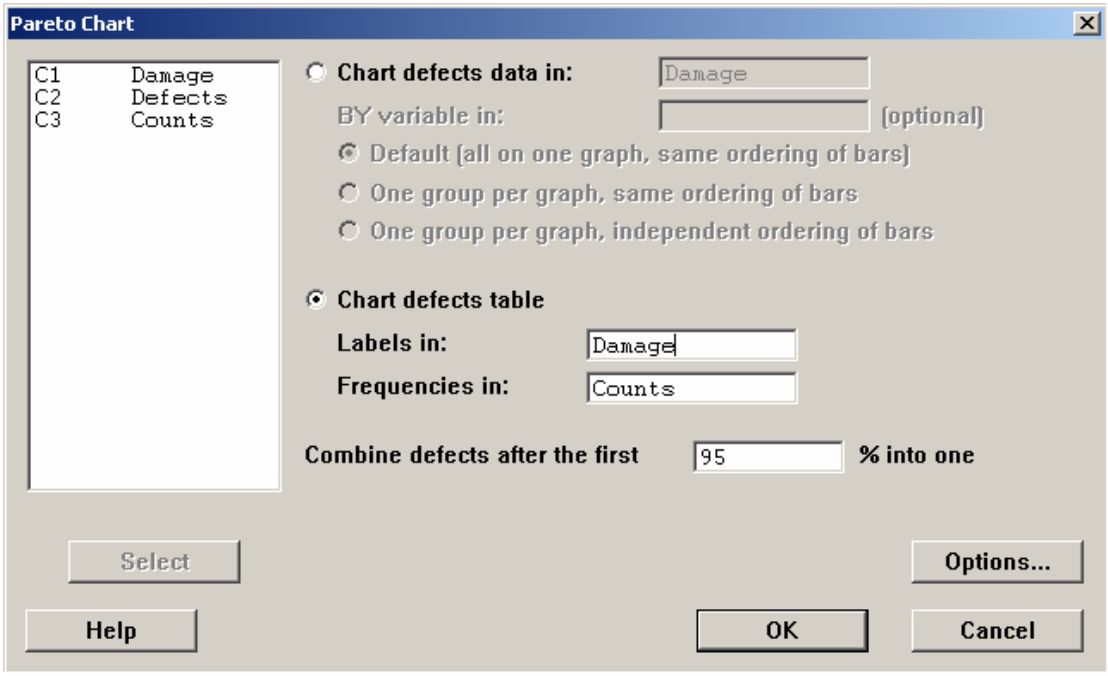
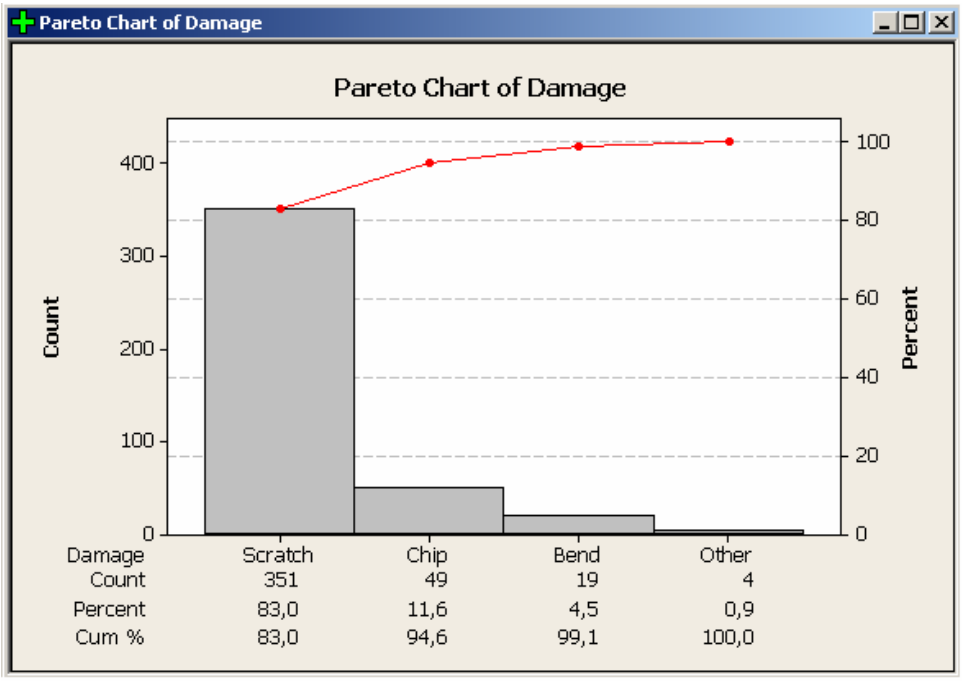


Chart defects table: Los defectos se han tabulado. En una columna aparecen los nombres (una sola vez) y en otra la frecuencia de aparición de cada uno de ellos. En este caso:

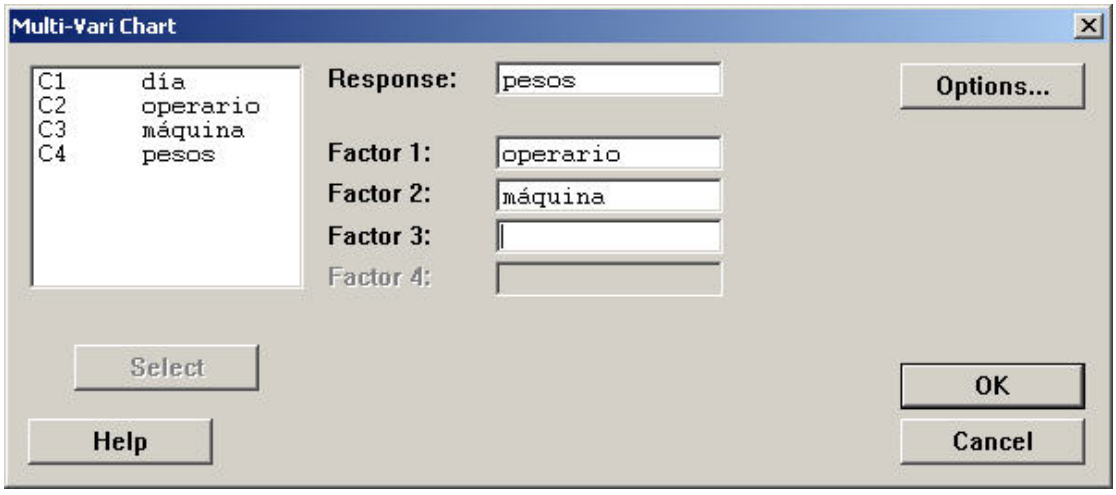


El resultado es:



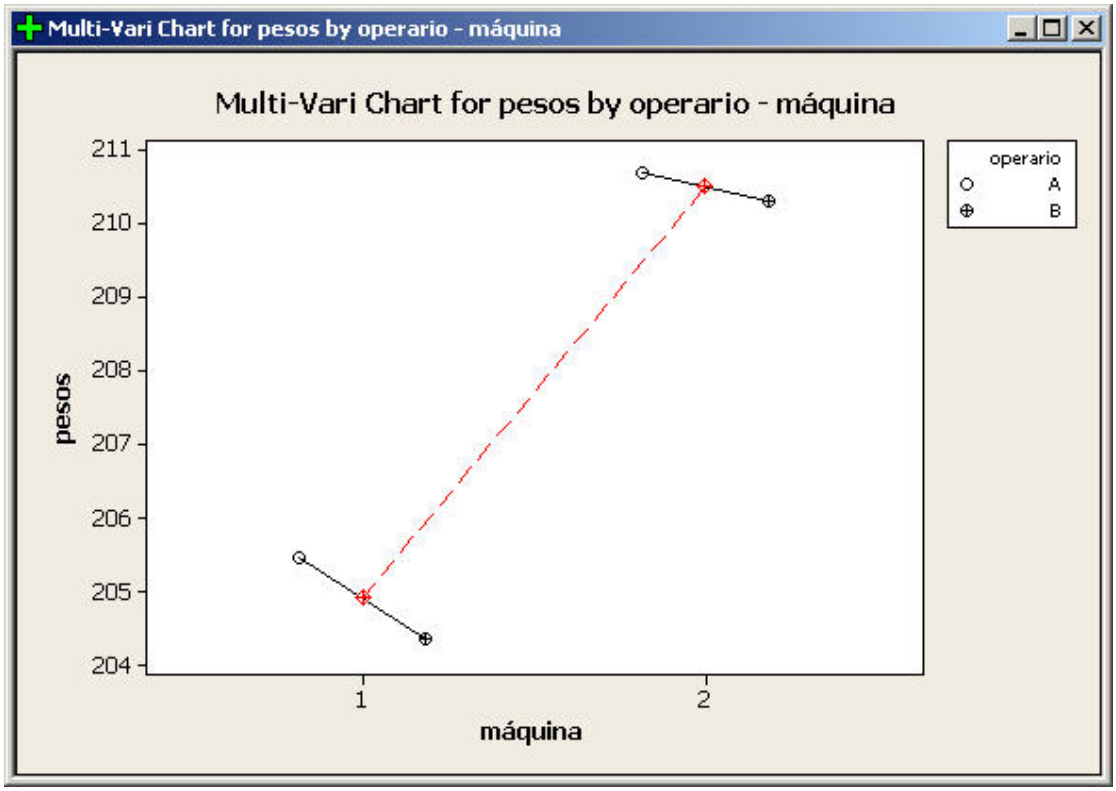
2.3.7 Multi-Vari

Es una forma muy visual de presentar la variabilidad de los datos cuando esta puede venir provocada por distintos factores. Minitab dibuja gráficos multi-vari en función de hasta cuatro factores. Los gráficos muestran las medias de las respuestas en función de los factores.



Se accede con **Stat/Quality Tools/Multi-Vari Chart**

El aspecto del gráfico depende del orden en que se introducen las fuentes de variación.



El resultado, en este caso, es:

Las opciones permiten eliminar líneas de conexión para que el gráfico se vea más claro.

2.3.8 Características de los gráficos

Como se ha podido ver a lo largo de los ejemplos, existen unas opciones de personalización de los gráficos muy similares. Se citan aquí las características más importantes:

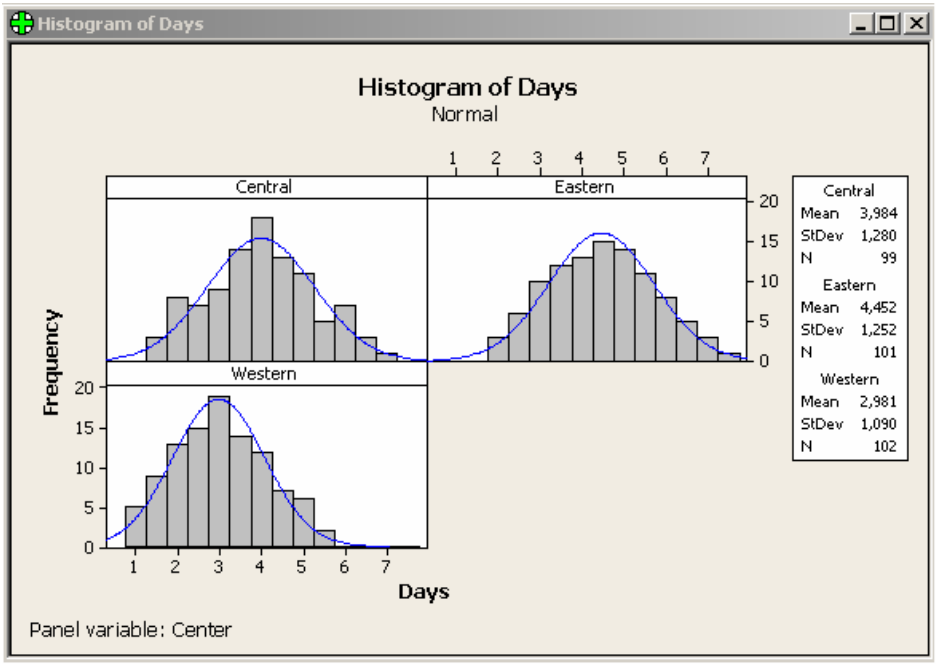
Scale: Esta opción permite cambiar y añadir los ejes, la cuadrícula y líneas de referencia así como otras características referentes a los ejes.

Labels: Para añadir o personalizar títulos, notas al pie de página y etiquetas.

Data View: Para añadir o personalizar símbolos que representan los datos en los gráficos, líneas de conexión de puntos, líneas de regresión, etc…

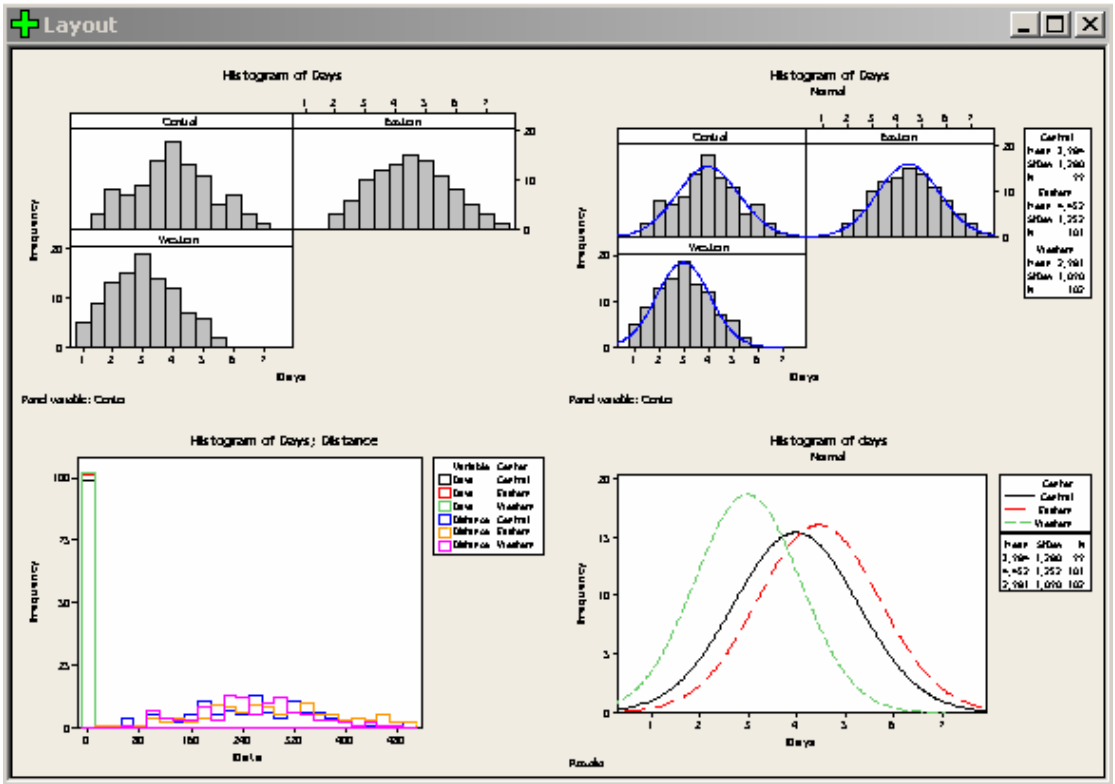
Multiple Graphs: Se pueden agrupar las variables utilizadas según alguna otra y, en función de esto, separar las diferentes gráficas en distintos paneles de un mismo dibujo. Para más información, consultar el Help.

Para lograr esto se clica en *By Variables* y se introduce en *By variables with groups on separate panels* la variable en función de la cual se quiere agrupar.



Una vez terminado el gráfico, se pueden modificar algunas de sus propiedades pulsando dos veces sobre el elemento a modificar.

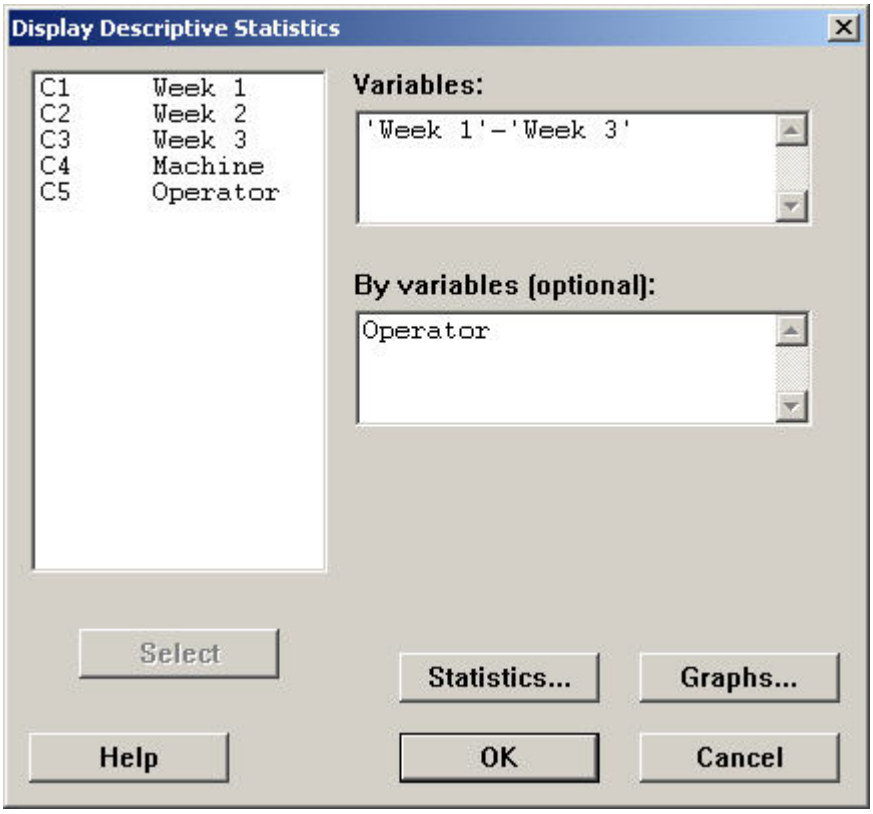
Por último se verá la opción **Editor/Layout Tool,** permite situar de forma muy intuitiva distintos gráficos realizados en una misma tabla.



3 ESTADÍSTICA DESCRIPTIVA

3.1 Estadística descriptiva

La ruta **Stat/Basis Statistics/Display Descriptive Statistics** permite obtener una tabla que contiene las características y funciones estadísticas más simples de las variables que se elijan. Pulsando sobre el botón Statistics se muestran los tipos disponibles. Se detallan los que más se utilizarán en la asignatura de estadística:



1. <u>Índices de posición (valores percentiles)</u>

First quartile: Calcula el valor que deja a la izquierda el 25% de los valores (Q_1)

Median: Calcula el valor que deja a la izquierda el 50% de los valores (Q_2 o mediana)

Third quartile: Calcula el valor que deja a la izquierda el 75% de los valores (Q_1)

Interquartile range: Indica la diferencia entre el primer y tercer cuartil.

2. <u>Índices de tendencia central</u>

Mean: Media aritmética.

Median: Calcula el valor que deja a la izquierda el 50% de los valores (Q_2 o mediana)

Sum: Suma de todos los valores

3. Índices de dispersión (Dispersión)

SE of mean: Error tipo de la media. Estimación de la variabilidad muestral de la media.

Standard deviation: Variabilidad de los valores con respecto a la media, expresada en las mismas unidades que los datos.

Variance: Variabilidad de los valores con respecto a la media, expresada en unidades al cuadrado.

Minimum: Valor más pequeño.

Maximum: Valor más grande.

Range: Diferencia entre los valores máximo y mínimo.

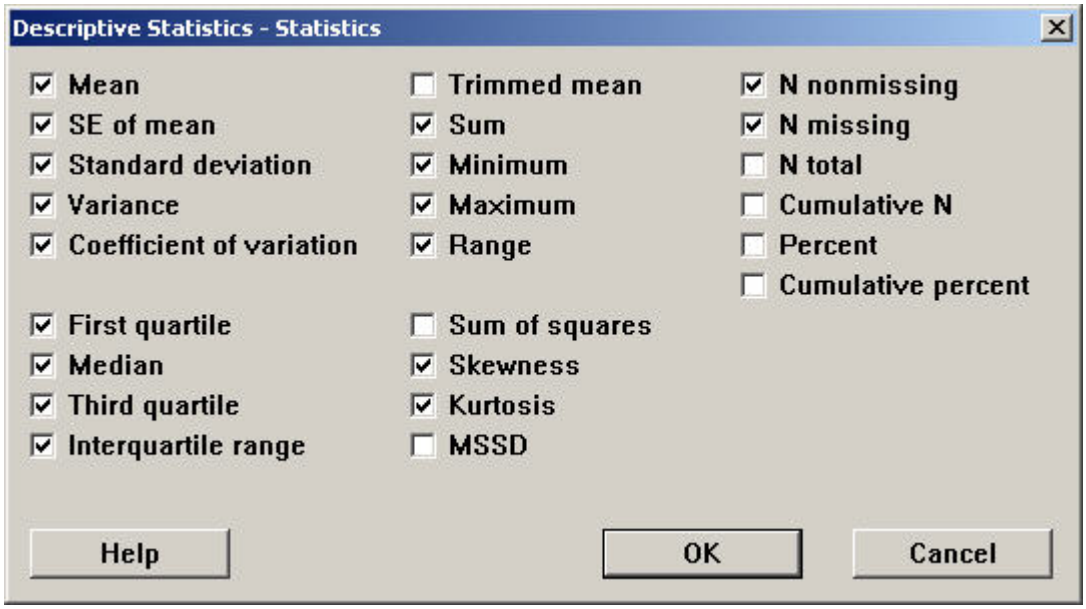
N nonmissing: Número de entradas que no faltan.

N missing: Número de entradas que faltan, (*).

4. Índices de distribución

Skewness: Coeficiente de asimetría.

Kurtosis: Coeficiente de curtosis.



El resultado aparece en una tabla en la ventana Session. No se pueden modificar las propiedades de la tabla, sólo su tipo de letra y similares. Por eso será importante elegir correctamente las variables y cuidar su disposición.

Como puede comprobarse, Minitab permite dibujar los gráficos más habituales desde la ventana de datos estadísticos.

4 REGRESIÓN Y CORRELACIÓN

Estas técnicas estadísticas sirven para comprobar la existencia o no de relación entre dos variables y, en el supuesto de que exista, ver cómo es esta relación: si es lineal, exponencial, cúbica… Como la relación nunca va a ser perfecta (lo sería si las variables fueran deterministas, no aleatorias como es el caso), es importante también en estos estudios cuantificar el grado de la relación supuesta y, para ello, se utiliza el el coeficiente de correlación. Este coeficiente que oscila entre -1 y 1, valorando con 0 la no relación y con 1 o -1 la relación perfecta positiva o negativa, respectivamente.

4.1 Correlación

En **Stat/Basic Statistics/Correlation** aparece un cuadro de diálogo donde pueden elegirse las variables para las cuales se quiere calcular el coeficiente de correlación lineal (es decir, aquel que cuantifica el grado de relación entre las variables a través de una recta) y el p-valor (valor que permite decidir si la recta puede ser un modelo de relación "significativo" (por debajo del 0,05) para las variables consideradas o no). Si se eligen más de dos variables, el MINITAB obtiene los coeficientes de correlación 2 a 2 y los representa matricialmente.

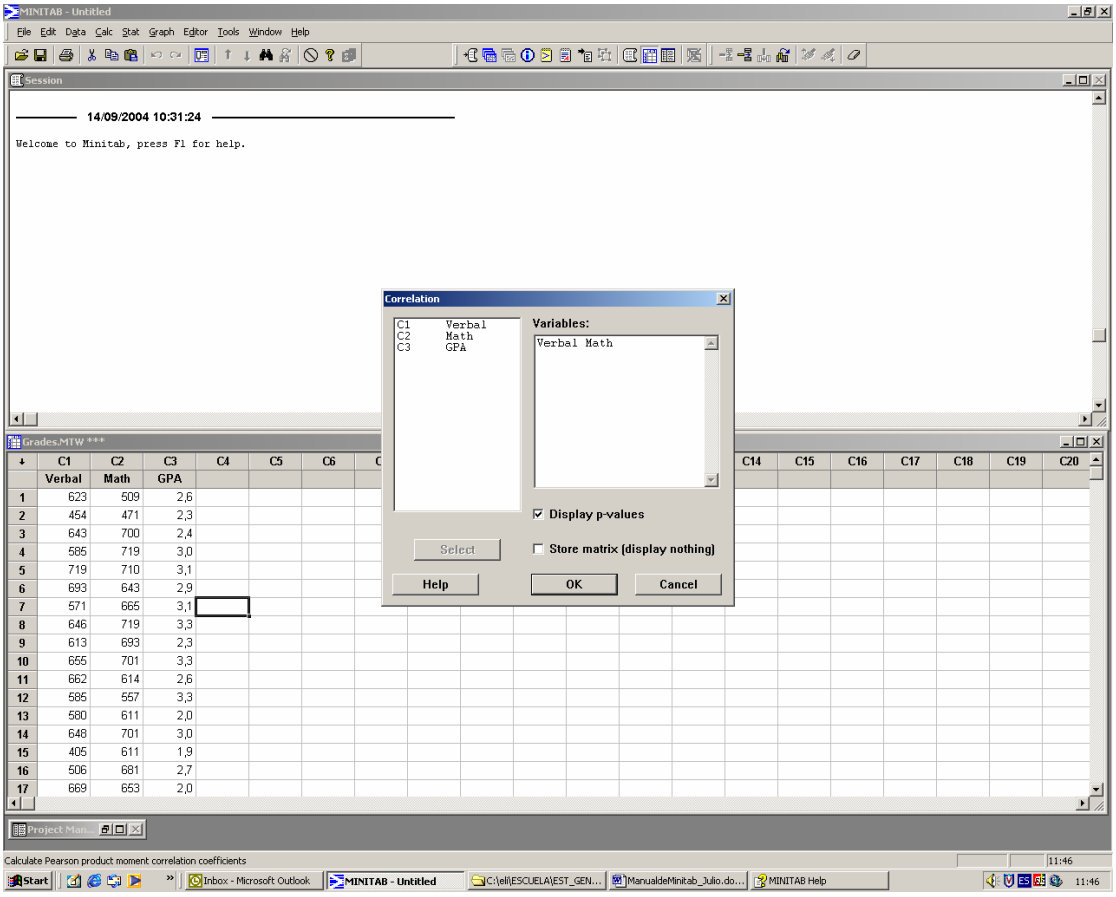
Por ejemplo, utilizando los datos referidos a puntuaciones en lengua (verbal) y matemáticas (math) de 200 alumnos, se ha calculado el coeficiente de correlación lineal entre ambas variables.

El resultado fue:

```
Pearson correlation of Verbal and Math = 0,275
P-Value = 0,000
```

Es decir, la relación lineal entre ambas variables es claramente significativa.

NOTA: Es importante entender que una correlación significativa no depende solo de su valor, sino también del tamaño de las muestras a partir de la cual se ha calculado. Por eso nunca se deben comparar coeficientes de correlación calculados de conjuntos de datos de distintos tamaños.



	C1	C2	C3
	Verbal	Math	GPA
1	623	509	2,6
2	454	471	2,3
3	643	700	2,4
4	585	719	3,0
5	719	710	3,1
6	693	643	2,9
7	571	665	3,1
8	646	719	3,3
9	613	693	2,3
10	655	701	3,3
11	662	614	2,6
12	585	557	3,3
13	580	611	2,0
14	648	701	3,0
15	405	611	1,9
16	506	681	2,7
17	669	653	2,0

Graph/Scatterplot da la opción de representar los datos para ver gráficamente la posible relación entre las variables (ver pág. 22)

4.2 Regresión Lineal Simple

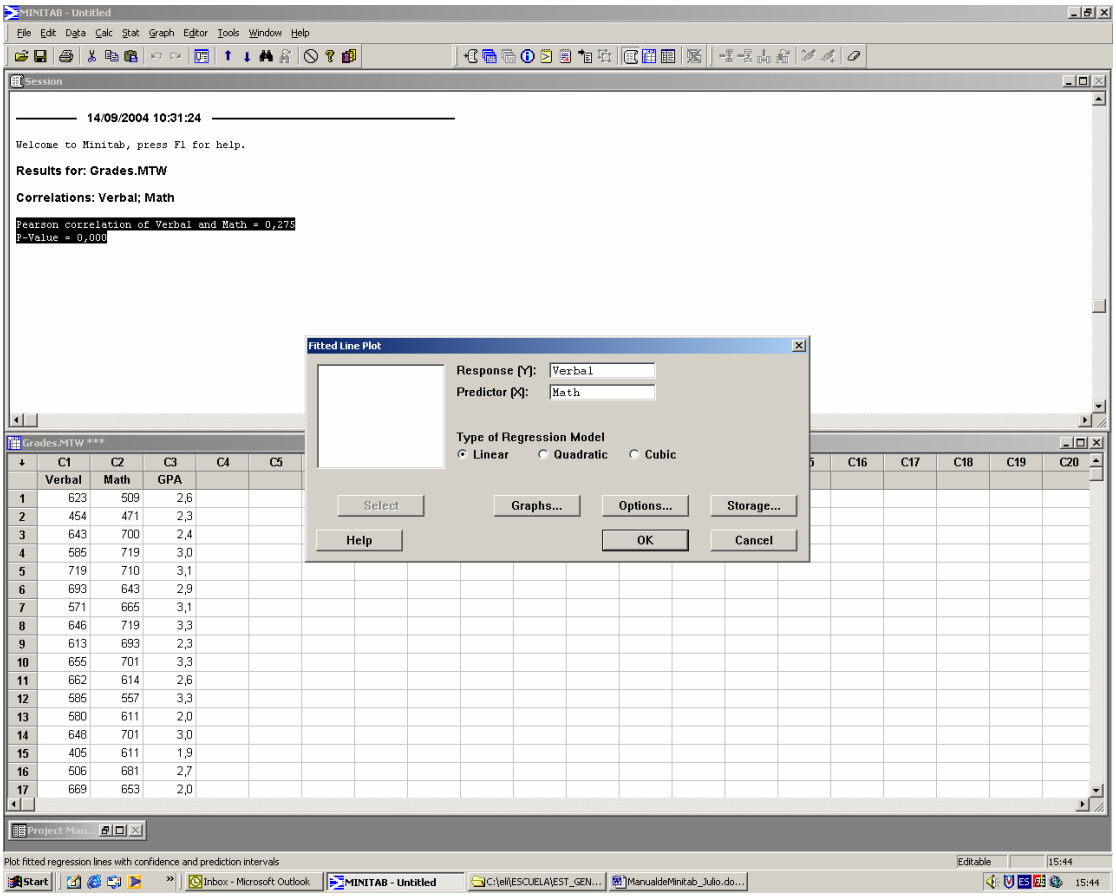
El procedimiento de ajustar los datos a un modelo de regresión lineal simple:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

se puede realizar desde dos opciones, que veremos a continuación:

- **Stat/Regression/Regression**: donde MINITAB proporciona una información sobre el análisis de rgeresión muy detallado a través de la ventana de Session.
- **Stat/Regression/Ftted Line Plot**: donde MNINITAB resuelve el problema de manera menos detallada, pero muestra una diagrama de dispersión de los datos, que completa gráficamente la información aportada.

4.2.1 Stat/Regression/Ftted Line Plot



	C1	C2	C3	C4	C5
	Verbal	Math	GPA		
1	623	509	2,6		
2	454	471	2,3		
3	643	700	2,4		
4	585	719	3,0		
5	719	710	3,1		
6	693	643	2,9		
7	571	665	3,1		
8	646	719	3,3		
9	613	693	2,3		
10	655	701	3,3		
11	662	614	2,6		
12	585	557	3,3		
13	580	611	2,0		
14	648	701	3,0		
15	405	611	1,9		
16	506	681	2,7		
17	669	653	2,0		

En esta opción se abre una ventana de diálogo, donde se pide que se inserte la variable respuesta Y y la predictora X. Se puede realizar un ajuste lineal, cuadrático o cúbico y con todas las opciones por defecto, se obtiene en la ventana Session lo siguiente:

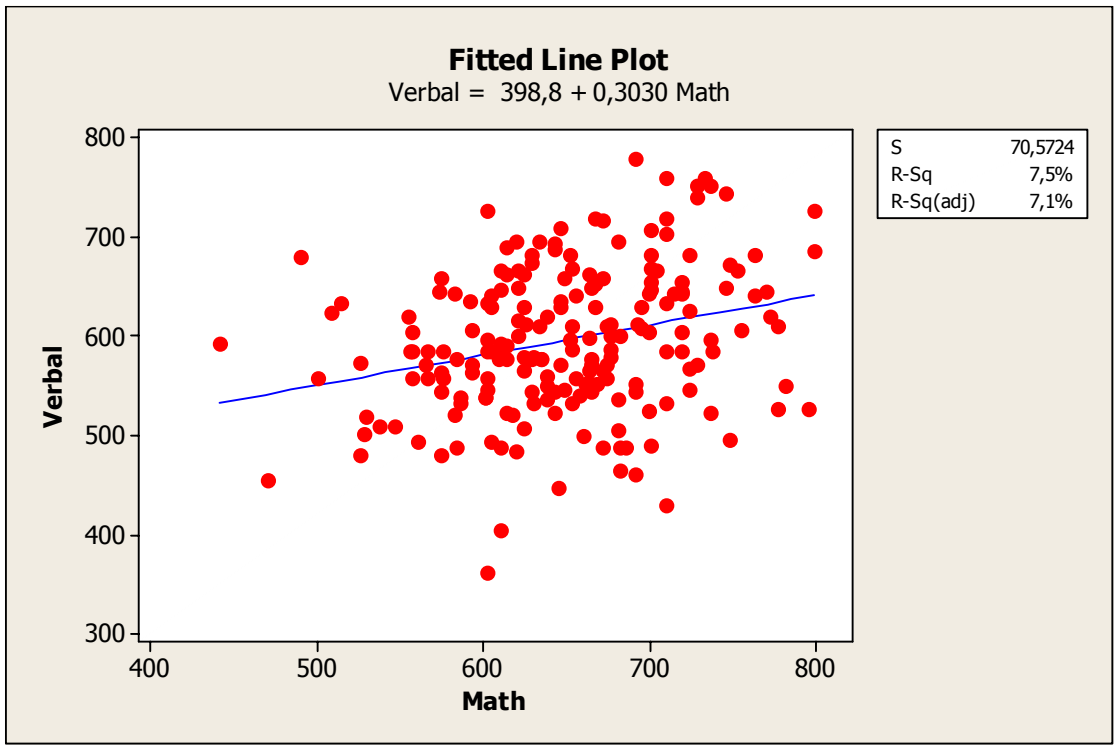
Regression Analysis: Verbal versus Math

```
The regression equation is
Verbal = 398,8 + 0,3030 Math


S = 70,5724   R-Sq = 7,5%   R-Sq(adj) = 7,1%


Analysis of Variance

Source        DF       SS        MS      F      P
Regression     1    80445   80445,4  16,15  0,000
Error        198   986132    4980,5
Total        199  1066577
```

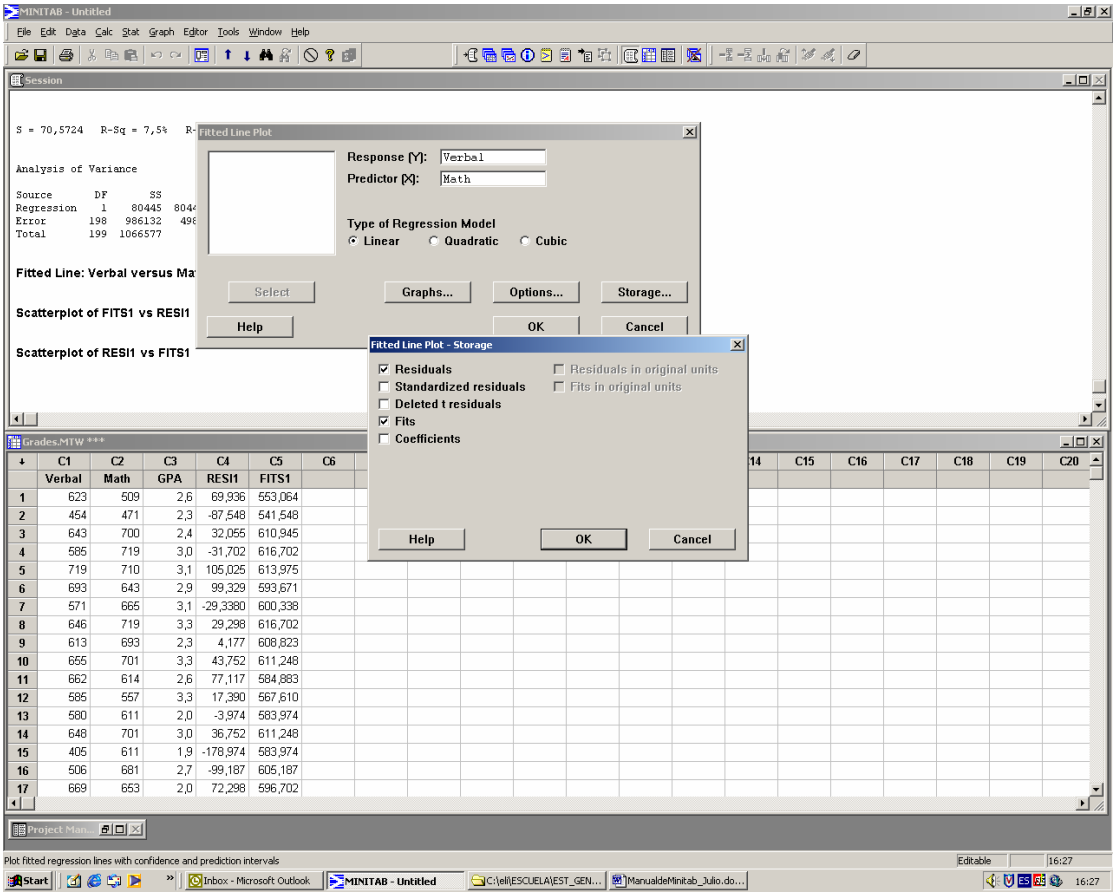
Se trata de la resolución de la regresión, mediante un análisis de varianza, donde el p-valor (en este ejemplo igual a 0,000 < 0,005) permite revelar que la regresión lineal es estadísticamente significativa. Además, en primer lugar aparece la recta de regresión ajustada y datos referentes a la desviación típica S de los residuos

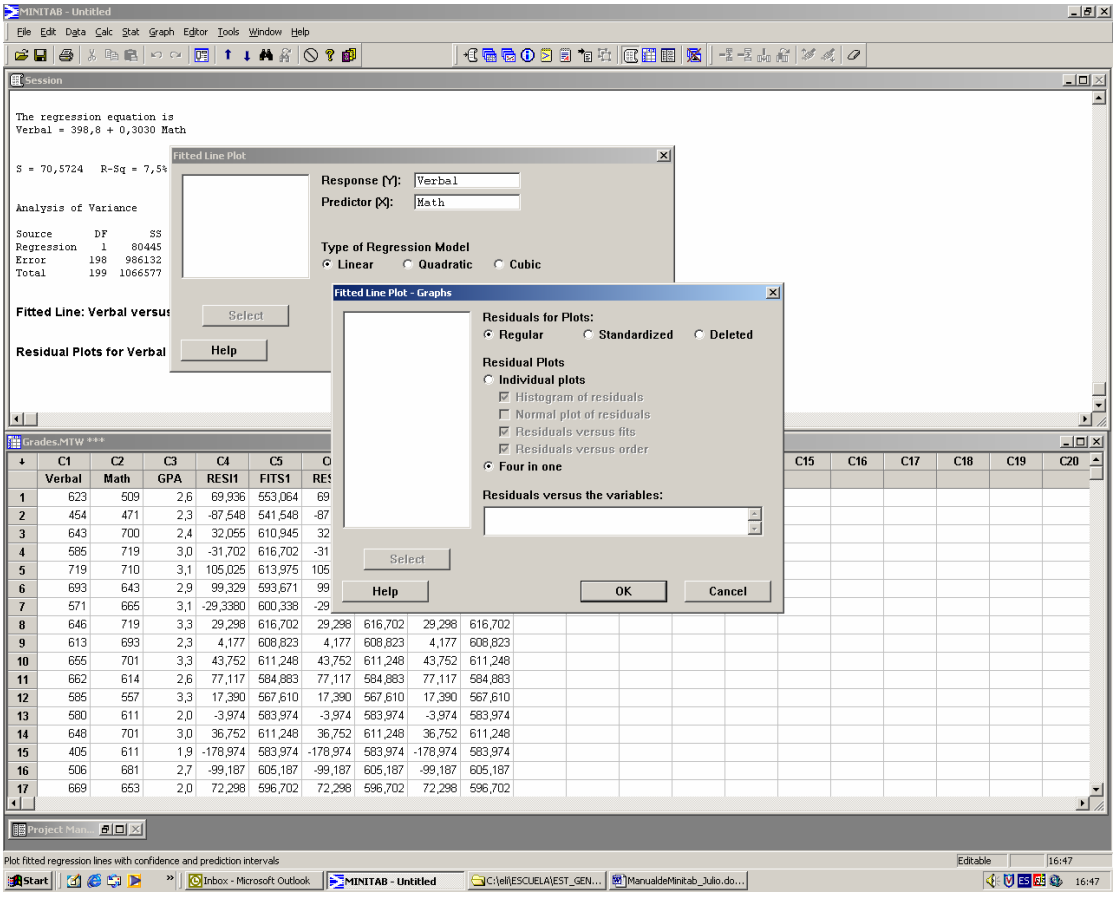
(residuo=valor real de respuesta-valor ajustado por la recta), R-Sq, que también se llama coeficiente de determinación y que es el resultado de elevar al cuadrado el coeficiente de correlación (se da en %) y permite medir la calidad del ajuste y, R-Sq (adj) que es el coeficiente de determinación ajustado (sólo tiene sentido en la regresión múltiple)

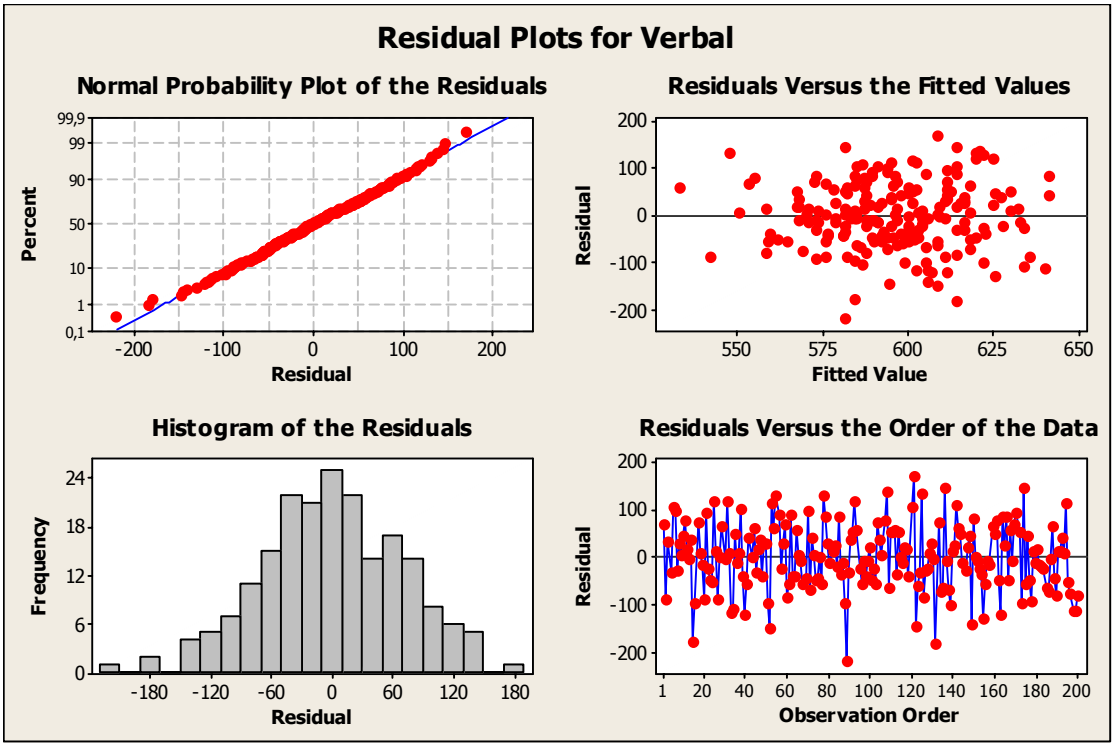
Además, el resultado se acompaña de una gráfica como esta, que resume lo anterior (salvo el valor del p-valor):



Para estar seguros de que se ha realizado un buen análisis de la posible relación entre las variables, (en el sentido de que se ha podido extraer de los datos toda la posible información contenida en ellos respecto de esa relación y, por tanto, esta solución no es mejorable con los datos disponibles), conviene analizar los residuos y comprobar si se verifican todos las hipótesis que se requieren en la aplicación de esta técnica y, por tanto, los residuos se comportan de manera aleatoria, distribuidos de acuerdo a una distribución normal, con varianza constante. Esto puede realizarse guardando, previamente mediante el botón **'Storage'** los residuos y los valores previstos (fits) y luego se observa la normalidad de los residuos y se representa mediante un gráfico scatterplot los residuos frente a los 'fits'; o, directamente, con la opción **'Graphs'** que da la opción de recoger los 4 gráficos más importantes de la validación de manera agrupada:


MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
S = 70,5724 R-Sq = 7,5% R-
Analysis of Variance
Source DF SS
Regression 1 80445 804
Error 198 986132 49
Total 199 1066577
Fitted Line: Verbal versus Ma
Scatterplot of FITS1 vs RESI1
Scatterplot of RESI1 vs FITS1
Fitted Line Plot
Response (Y): Verbal
Predictor (X): Math
Type of Regression Model
Linear Quadratic Cubic
Select
Graphs...
Options...
Storage...
Help
OK
Cancel
Fitted Line Plot - Storage
Residuals
Standardized residuals
Deleted t residuals
Fits
Coefficients
Residuals in original units
Fits in original units
Help
OK
Cancel
Grades.MTW ***
C1 C2 C3 C4 C5 C6 C15 C16 C17 C18 C19 C20
Verbal Math GPA RESI1 FITS1
1 623 509 2,6 69,936 553,064
2 454 471 2,3 -87,548 541,548
3 643 700 2,4 32,055 610,945
4 585 719 3,0 -31,702 616,702
5 719 710 3,1 105,025 613,975
6 693 643 2,9 99,329 593,671
7 571 665 3,1 -29,3380 600,338
8 646 719 3,3 29,298 616,702
9 613 693 2,3 4,177 608,823
10 655 701 3,3 43,752 611,248
11 662 614 2,6 77,117 584,883
12 585 557 3,3 17,390 567,610
13 580 611 2,0 -3,974 583,974
14 648 701 3,0 36,752 611,248
15 405 611 1,9 -178,974 583,974
16 506 681 2,7 -99,187 605,187
17 669 653 2,0 72,298 596,702
Project Man...
Plot fitted regression lines with confidence and prediction intervals
Editable
16:27
Start
Inbox - Microsoft Outlook
MINITAB - Untitled
C:\eli\ESCUELA\EST_GEN...
ManualdeMinitab_Julio.do...
16:27





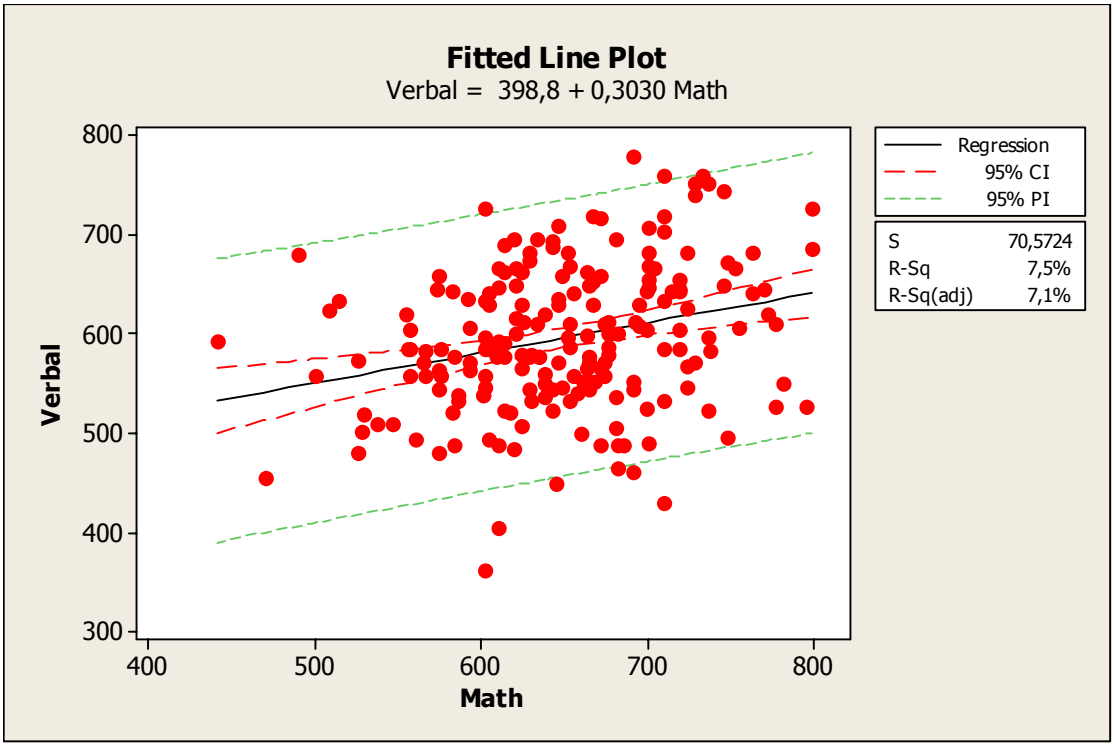
Si en el gráfico de los residuos versus fits, los puntos aparecen repartidos de manera aleatoria alrededor del valor 0 y en residuales frente al orden no se observa

tendencia alguna, quedará validado el análisis de regresión. La no normalidad de los residuos o la observación de alguna forma curvilínea en el gráfico residuales-fits sugerirá el ajuste de los datos a otra curva.

Una vez validada la curva de ajuste, su utilización práctica será la predicción. Se podrá obtener una estimación del valor de Y para un x_i particular desde el botón **'options'**. Una vez presionado aparece un nuevo cuadro de diálogo donde se puede decir que en el gráfico de dispersión donde se dibuja la curva de ajuste, se dibujen también las bandas de confianza para el valor medio y para valores individuales con un determinado nivel de confianza (usualmente 95%).

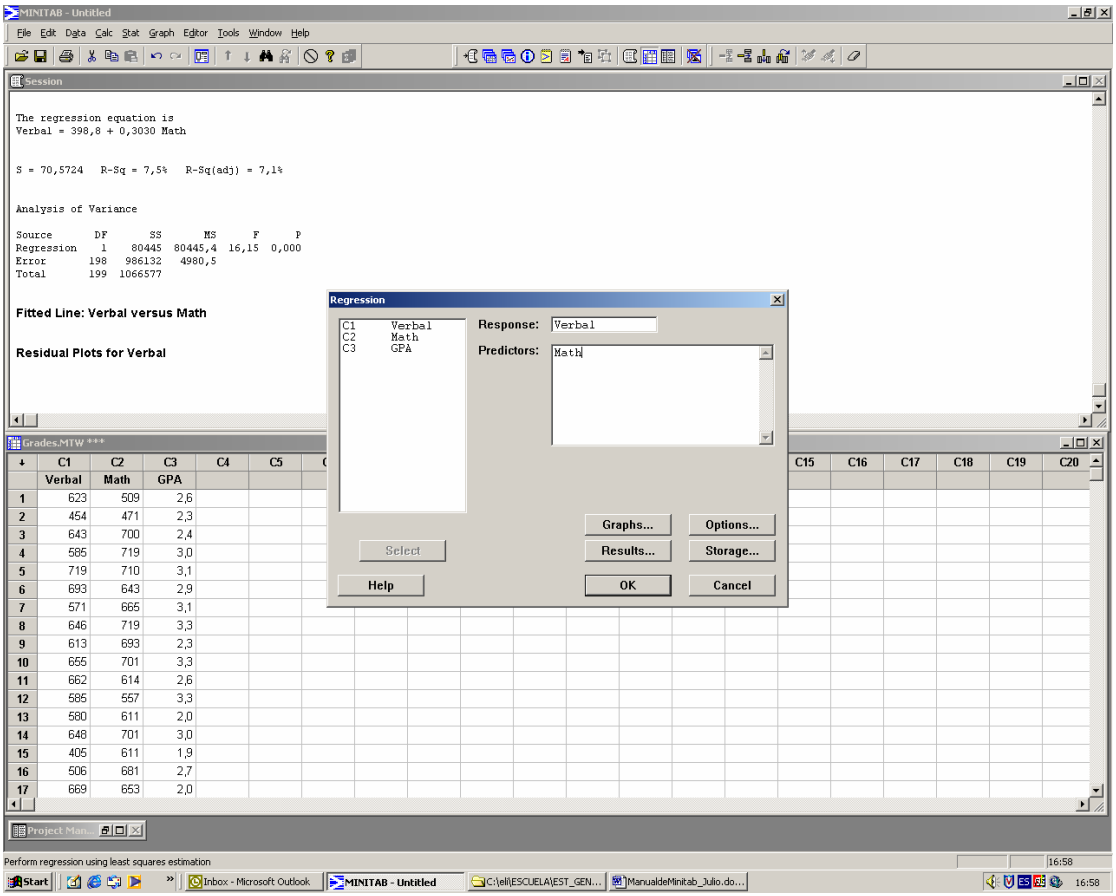
El intervalo de confianza para valores medios (CI) , representada por la banda más estrecha (la más cercana a la curva de ajuste), especifica entre qué valores se estima que está el promedio de los posibles valores de Y para un x_i determinado.

El intervalo de confianza para valores individuales (PI), representado por la banda más ancha, especifica entre qué valores se estima que se encuentre una única observación individual Y para un x_i determinado.



4.2.2 Stat/Regression/Regression

Como se ha comentado anteriormente, otra posibilidad es realizar el análisis de regresión lineal a través de esta ruta. En este caso el cuadro de diálogo es el siguiente:



Existe un hueco para elegir una única variable respuesta, pero hay posibilidades para elegir más de una variable predictora. En cualquier caso, eligiendo de nuevo estudiar la posible relación entre las puntuaciones de lengua y las de matemáticas y dejando todas las opciones por defecto, esto es lo que aparece en Session:

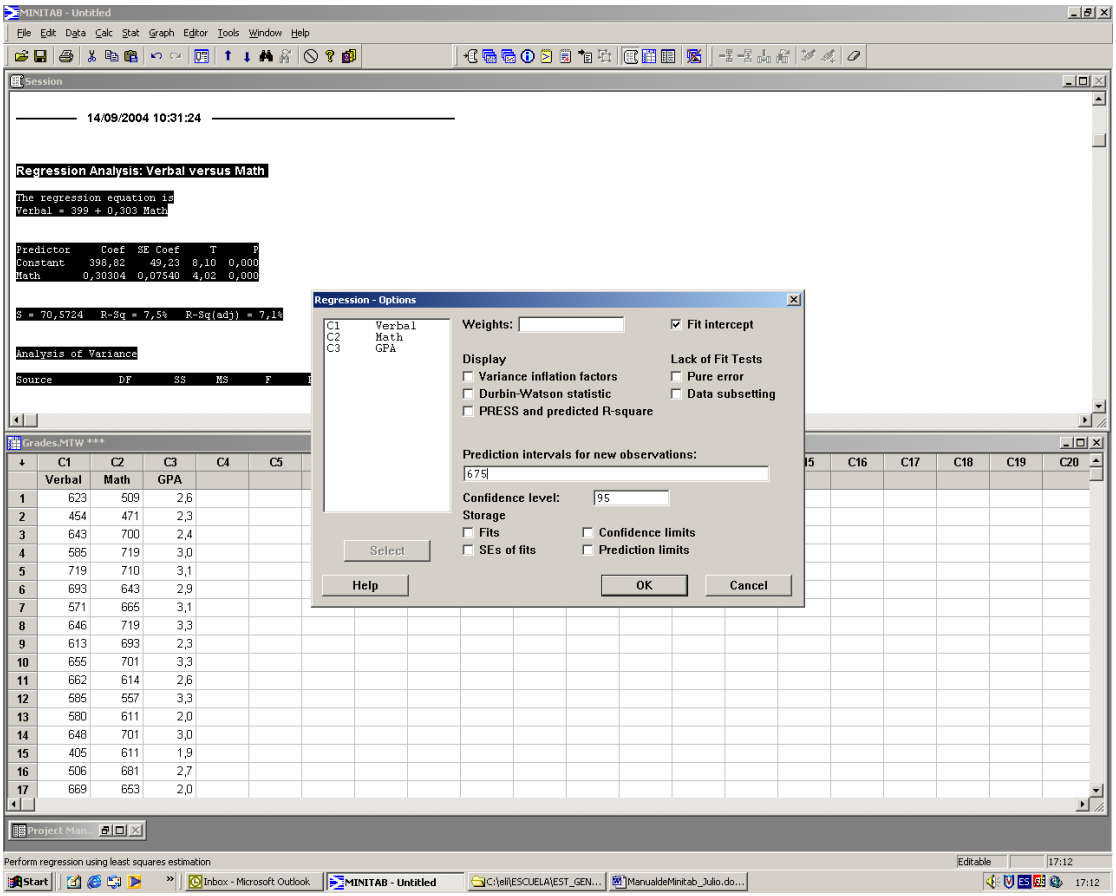
Regression Analysis: Verbal versus Math

```
The regression equation is
Verbal = 399 + 0,303 Math


Predictor      Coef  SE Coef      T      P
Constant     398,82    49,23   8,10  0,000
Math        0,30304  0,07540   4,02  0,000


S = 70,5724   R-Sq = 7,5%   R-Sq(adj) = 7,1%

Analysis of Variance

Source           DF       SS     MS      F      P
Regression        1    80445  80445  16,15  0,000
Residual Error  198   986132   4980
Total           199  1066577
```

```
Unusual Observations

Obs   Math   Verbal      Fit   SE Fit   Residual   St Resid
  2    471   454,00   541,55    14,36     -87,55      -1,27 X
 15    611   405,00   583,97     5,77    -178,97      -2,54R
 18    500   558,00   550,34    12,33       7,66       0,11 X
```

Aparece en primer lugar la ecuación de la recta de ajuste, con los valores de los coeficientes y sus p-valores para decidir si son o no significativamente distintos de cero, la desviación típica de los residuos y los coeficientes de determinación del ajuste, el análisis de varianza con el p.—valor, que permite decidir si la regresión lineal es significativa y, por último, observaciones "raras" en el sentido de que tienen residuos muy grandes (aparecen marcados con una R) o puntos muy influyentes sobre la recta (marcados con una X). A estos puntos conviene prestarles una atención mayor para valorar la conveniencia de mantenerlos en el estudio.

Los botones del cuadro de diálogo que amplian los resultados que se pueden obtener del estudio son similares a los que aparecen en la ruta anterior. Cabe destacar que en el caso del botón 'options' se permite incluir valores de x_i para estimar sus predicciones:



dando como resultado en el caso del ejemplo:

```
Predicted Values for New Observations

New
Obs      Fit  SE Fit        95% CI             95% PI
  1  603,37    5,35  (592,82; 613,91)  (463,80; 742,94)
```

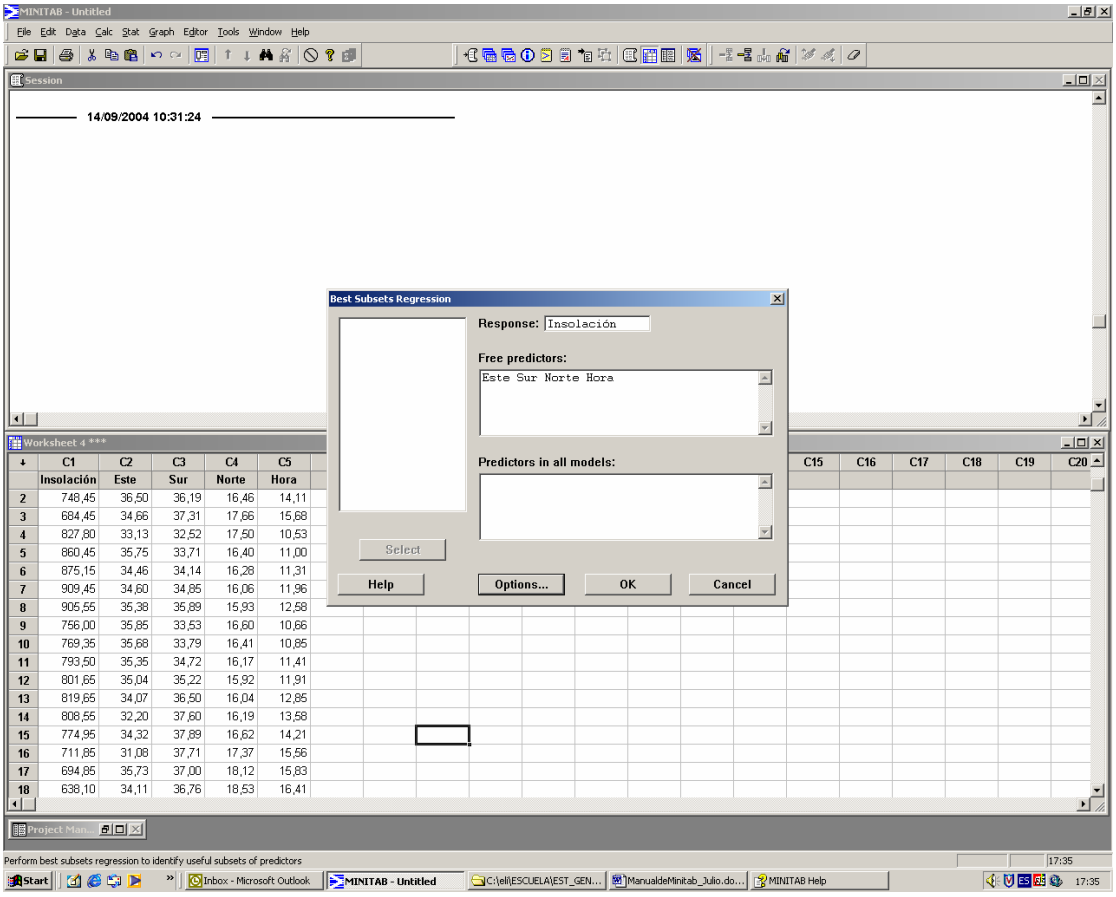
El uso de esta ruta para la realización de un estudio de regresión lineal se justifica, sobre todo, cuando existen más de una posible variable predictora. Es lo que se conoce como regresión múltiple. En estos casos, el problema real que suele plantearse es que, de entrada, uno no sabe qué variables van a explicar mejor la respuesta (con que variables X's se relaciona mejor Y). MINITAB resuelve este problema de dos maneras distintas:

4.2.3 Selección de la mejor ecuación: Stat/Regression/Best Subsets

Este método genera todas los modelos de regresión posibles a partir de una variable respuesta una serie de variables predictoras y valores de cuantificación del ajuste que permita realizar la elección de la que parezca más conveniente estudiar a fondo. Veámoslo con un ejemplo.

Se trata de intentar establecer la relación que existe entre el valor de insolación medido y variables relacionadas con la posición del punto de medición (Norte, este y sur) y el tiempo en el que se ha tomado la medida (hora).

El botón 'Options' permite establecer con cuántas variables predictoras, como máximo y como mínimo, va a formular los modelos (en nuestro caso como mínimo que elija la mejor X, y como máximo de la ecuación considerando las 4 variables X's introducidas). También permite establecer cuantas ecuaciones va a proponer para cada grupo de k variables predictoras (por defecto, propone las 2 mejores ecuaciones desde el modelo más simple, 1 X y 1 Y, hasta el más completo, todas las X's introducidas y la Y)



El resultado de esta opción aparece en Sesion como sigue:

Best Subsets Regression: Insolación versus Este; Sur; Norte; Hora

```
Response is Insolación

                                          N
                                   E      o H
                                   s S    r o
                          Mallows  t u    t r
Vars   R-Sq  R-Sq(adj)      C-p       S  e r    e a
   1   40,2      38,0       8,3  62,826       X
   1   34,1      31,6      11,7  65,983         X
   2   51,5      47,8       4,0  57,650    X    X
   2   48,1      44,1       5,9  59,676 X       X
   3   54,9      49,5       4,1  56,728    X  X X
   3   52,2      46,5       5,6  58,366 X  X    X
   4   56,9      49,7       5,0  56,599 X  X  X X
```

Y la interpretación es la siguiente:

Cada línea de la salida representa un modelo diferente. Vars hace referencia al número de variables predictoras en ese modelo, las cuales vienen indicadas con una X, al final. La bondad del ajuste se mide con el valor de R-Sq (adj.), que es el coeficiente de determinación ajustado al número de variables predictoras tomado (en

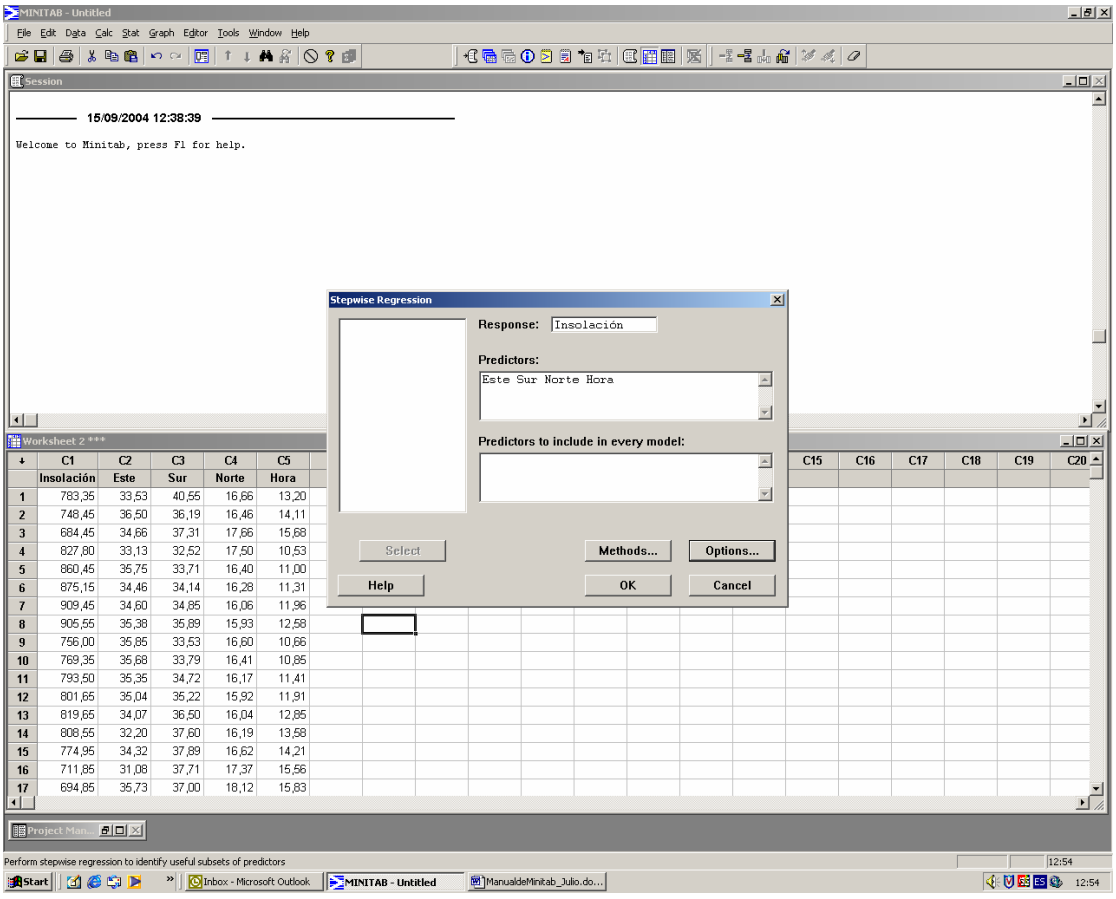
la regresión múltiple el R-sq, no tiene sentido) y cuanto más alto sea este porcentaje mejor. La s se refiere a la desviación de los residuos, luego cuanto menor sea, mejor.

De acuerdo entonces a estos criterios, en el ejemplo anterior podríamos decir que el mejor modelo (aquel que combina, facilidad de interpretación y buen ajuste) podría estar entre el que relaciona la insolación con la información posicional del Sur y la hora en que se toma el sol (utilizando solo dos variables regresoras) o el que relaciona la medida de insolación con los datos de Sur, Norte y Hora, es decir, añadiendo ola información de una variable más al modelo anterior.

Se trataría de elegir entre ambos modelos y dar una solución más completa para el elegido, adjuntando gráficos, la ecuación del modelo, su validación,...utilizando la opción Stat/Regression/Regression.

4.2.4 Selección de la mejor ecuación: Stat/Regression/Stepwise

Como el mismo nombre indica, en este caso, la selección del modelo se realiza paso a paso. En este caso, la ventana del menú pide introducir la variable respuesta y las variables predictoras candidatas a entrar en el modelo (por defecto y si no se tiene conocimiento previo, debieran ir todas; si hay algunas que se sabe que debieran estar en cualquier modelo creado, se puede también especificar).



Por defecto, el MINITAB va creando ecuaciones de regresión (modelos) añadiendo cada vez una variable más de la siguiente manera: comienza proponiendo un modelo con la variable predictora que mejor ajuste daría. El siguiente modelo propuesto añade una variable predictora al modelo anterior (la que da mejor resultado en el ajuste de la respuesta con dos variables predictoras), y así sucesivamente hasta que no observa ninguna variable más de entre las propuestas que añadan significación al modelo (para entrar una variable al modelo debe quedar dentro con un p-valor < 0,15; el mismo valor que para salir, si fuera necesario tras algún paso)

Se puede elegir como opción, botón **'Methods'** la posibilidad de crear modelos añadiendo variables (sin después salir) si el p-valor con el que quedan en la ecuación es menor que 0,25 por defecto (*Forward selection*) o partir del modelo más completo e ir sacando variables si su valor es mayor a 0,1, por defecto (*Backward elimination*).

En **'Options'** se permite mostrar otras opciones además de la que selecciona.

Con el método Stepwise y las opciones por defecto el resultado en Session para el ejemplo anterior sería el siguiente:

Stepwise Regression: Insolación versus Este; Sur; Norte; Hora

```
  Alpha-to-Enter: 0,15  Alpha-to-Remove: 0,15

Response is Insolación on 4 predictors, with N = 29

Step             1      2
Constant      1685   2291

Norte          -56    -59
T-Value      -4,26  -4,69
P-Value      0,000  0,000

Este                -15,9
T-Value             -1,98
P-Value             0,058

S             62,8   59,7
R-Sq         40,23  48,07
R-Sq(adj)    38,01  44,07
Mallows C-p    8,3    5,9
```

El primer modelo que presenta elige a la variable Norte (que de las 4 variables predictoras propuestas, es la que mayor información aporta sola, R^2=38%). El siguiente modelo lo deduce de elegir entre las tres variables predictoras restantes, alguna que aporte más información sobre le valor de la insolación. En este caso, resultaría de añadir la variable Este al modelo, lo que haría aumentar el valor de R^2 hasta el 48% , reduciendo algo el valor de S.

Si ya no entran más variables al modelo es porque su p-valor es superior a 0,15. Por tanto, se observa en los resultados, que la información que se proporciona es información sobre la/s variable/s que entraría/n al modelo, el coeficiente que acompañaría a las variables en la ecuación, y el valor del p-valor para ese coeficiente; también se acompaña la medida del ajuste del modelo (R^2) y la S. Para validar cualquiera de estos modelos, se debería realizar el elegido a través de la ruta Stat/Regression/Regression.

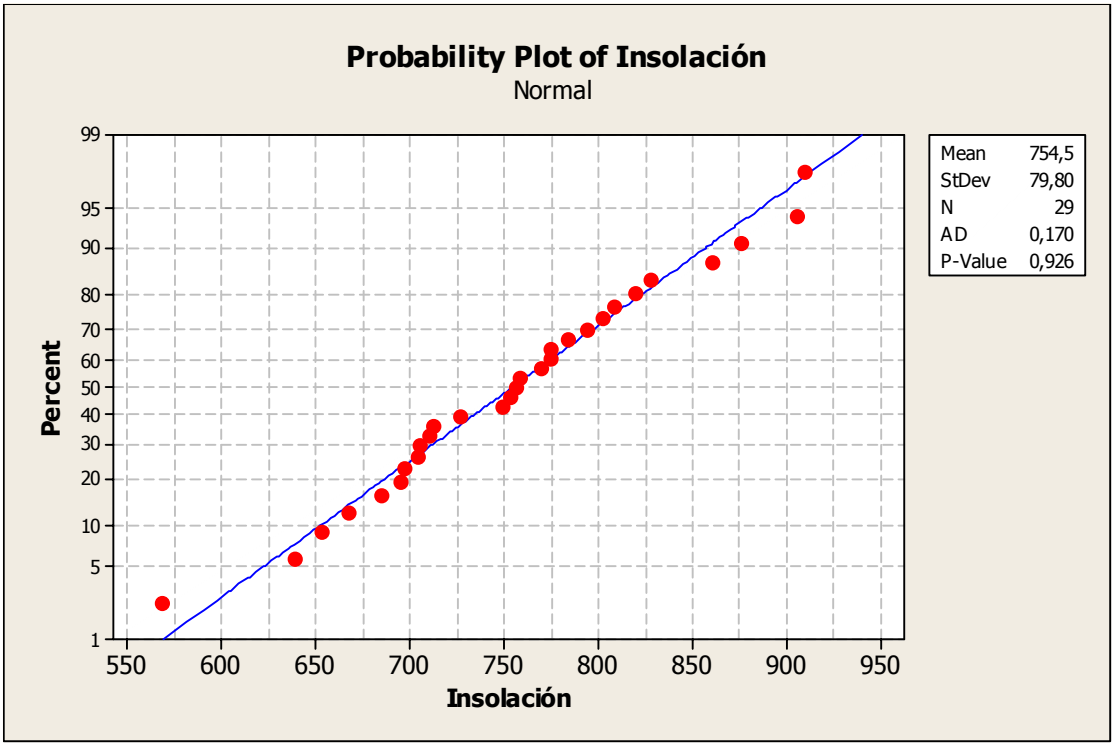
NOTA: Es interesante fijarse que el modelo de regresión que se propone por ambas vías de elección de la mejor ecuación, no tiene porque coincidir. Esto es lógico porque para realizar esta elección, no se ha seguido la misma estrategia.

5 TEST DE HIPÓTESIS

5.1 Test de normalidad

Resulta muy frecuente en la práctica el tener que comprobar si los datos con los que se está trabajando provienen de una distribución normal. Esto lo hace MINITAB desde **Stat/Basic Statistics/Normality Test**. En la ventana de diálogo que aparece en este caso, solo hay que indicarle en qué variable están los datos cuya normalidad se quiere comprobar y el programa permite elegir entre tres test de bondad de ajuste distintos para llevarlo a cabo. Por defecto está el test de Anderson-Darling, basado en comparaciones con la función de distribución, al igual que el de Kolmogorov-Smirnov y también se da como opción el método de Shapiro-Wilks, basado en la correlación. La respuesta en todos ellos va a tener la misma gráfica, la que representa el ajuste de los datos a PPN, lo cual se interpreta de la siguiente manera:

Cuanto más alineados estén los puntos, mejor. La calidad del ajuste se medirá de acuerdo al modelo. El p-valor en cada caso, indicará si se rechaza o no la hipótesis de normalidad (un p-valor por debajo de 0,05, estará revelando que los datos no son normales). Además se da como información la estimación de la media y de la desviación típica.

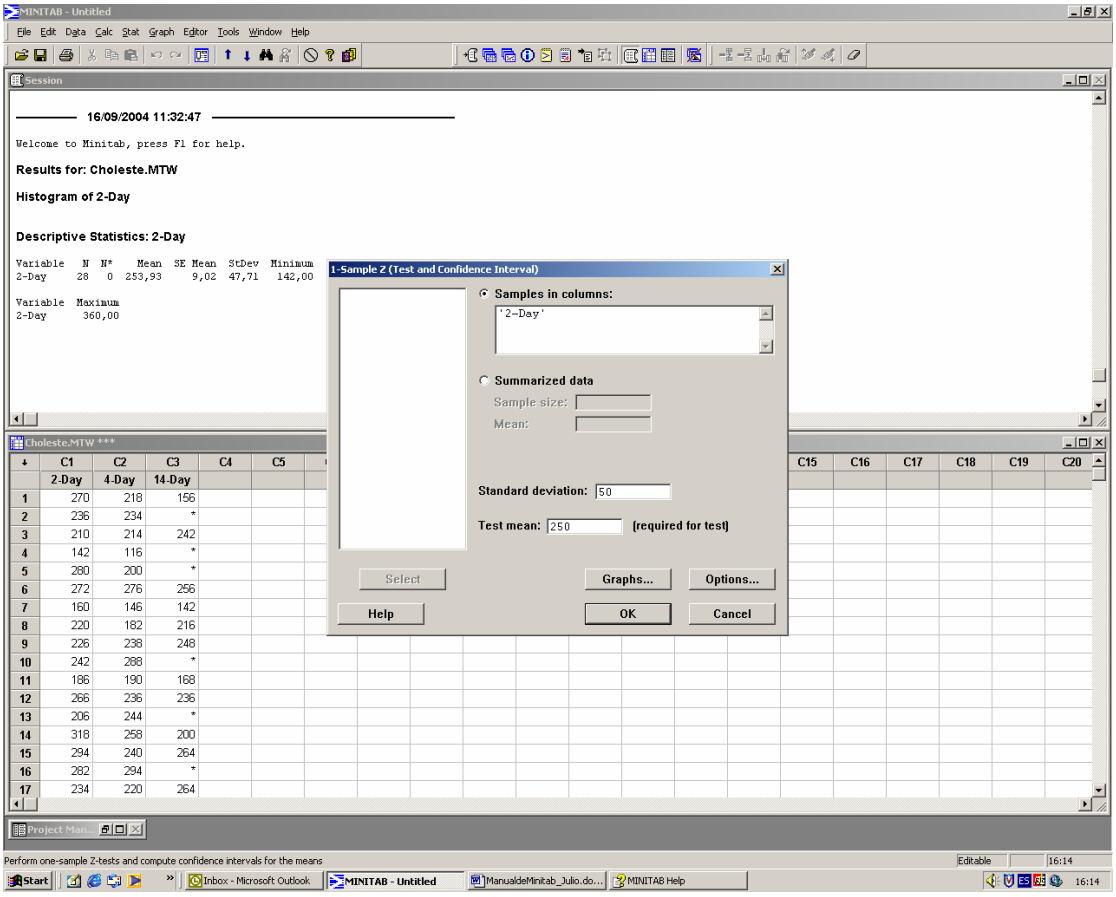


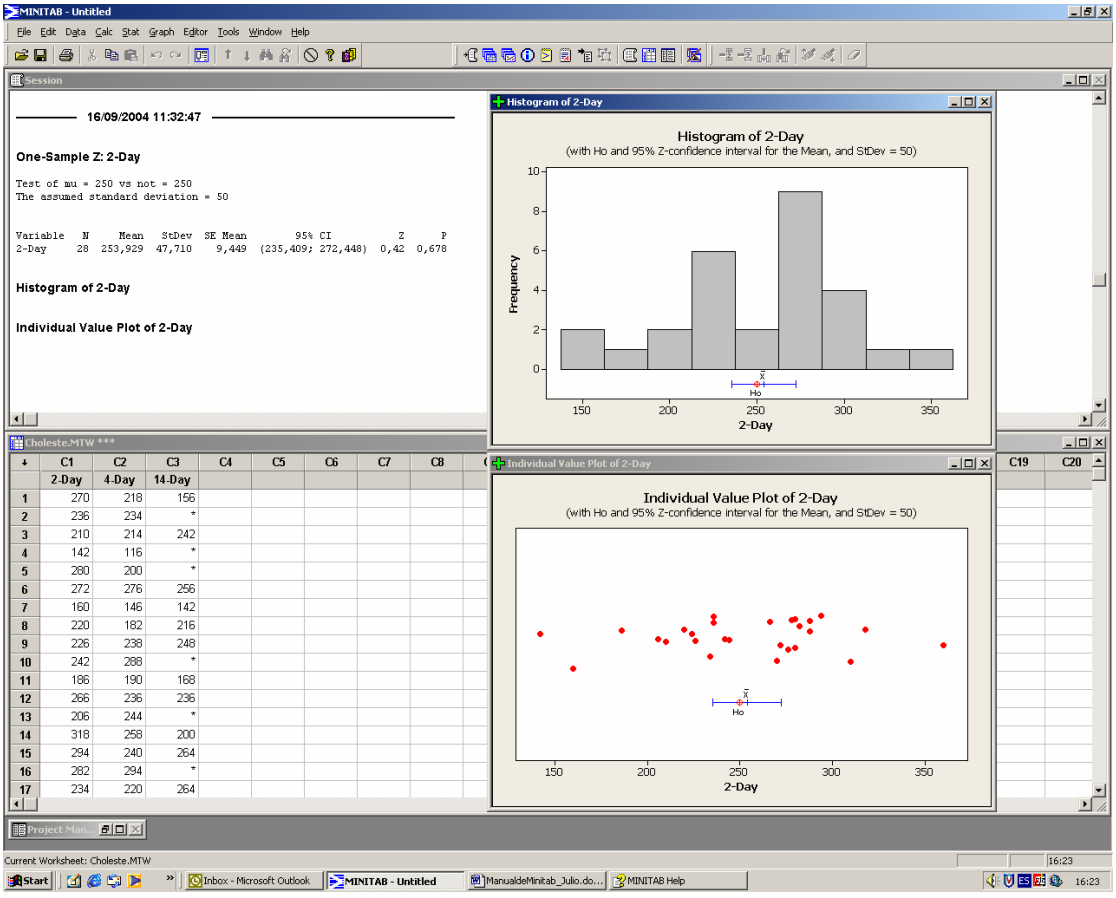
5.2 Intervalo de confianza y Test de hipótesis

MINITAB calcula los intervalos de confianza y resuelve los test de hipótesis más frecuentes de la siguiente manera:

5.2.1 Stat/Basic Statistics/1_sample Z

En esta prueba se trata de comprobar la hipótesis nula de que el valor medio de una variable, de desviación conocida, sigue una determinado valor. Es por esto que en la ventana de diálogo correspondiente se debe indicar la variable sobre la que se realizará el test, el valor de la desviación típica (que debe de ser conocido) y el valor medio para contrastar. Por defecto, si no se indica lo contrario, lo que se obtiene por respuesta es un intervalo bilateral y el resultado del test (el p-valor) El test que se realiza es bilateral (es decir, la hipótesis alternativa es que la media es distinta al valor especificado). Esto puede modificarse entrando por el botón **'Options'**. También es opcional poder representar los datos y esto se deberá especificar en **'Graphs'** eligiendo entre un histograma, un Dotplot o un BoxPlot.


MINITAB - Untitled
Session
16/09/2004 11:32:47
Welcome to Minitab, press F1 for help.
Results for: Choleste.MTW
Histogram of 2-Day
Descriptive Statistics: 2-Day
Variable N N* Mean SE Mean StDev Minimum
2-Day 28 0 253,93 9,02 47,71 142,00
Variable Maximum
2-Day 360,00
1-Sample Z (Test and Confidence Interval)
Samples in columns:
'2-Day'
Summarized data
Sample size:
Mean:
Standard deviation: 50
Test mean: 250 (required for test)
Select
Graphs...
Options...
Help
OK
Cancel
Choleste.MTW ***
C1 C2 C3 C4 C5 C15 C16 C17 C18 C19 C20
2-Day 4-Day 14-Day
1 270 218 156
2 236 234 *
3 210 214 242
4 142 116 *
5 280 200 *
6 272 276 256
7 160 146 142
8 220 182 216
9 226 238 248
10 242 288 *
11 186 190 168
12 266 236 236
13 206 244 *
14 318 258 200
15 294 240 264
16 282 294 *
17 234 220 264
Project Man...
Perform one-sample Z-tests and compute confidence intervals for the means
Editable
16:14
Start
Inbox - Microsoft Outlook
MINITAB - Untitled
ManualdeMinitab_Julio.do...
MINITAB Help
16:14


MINITAB - Untitled
Session
16/09/2004 11:32:47
One-Sample Z: 2-Day
Test of mu = 250 vs not = 250
The assumed standard deviation = 50
Variable N Mean StDev SE Mean 95% CI Z P
2-Day 28 253,929 47,710 9,449 (235,409; 272,448) 0,42 0,678
Histogram of 2-Day
Individual Value Plot of 2-Day
Histogram of 2-Day
(with Ho and 95% Z-confidence interval for the Mean, and StDev = 50)
Frequency
2-Day
Ho
Individual Value Plot of 2-Day
(with Ho and 95% Z-confidence interval for the Mean, and StDev = 50)
Choleste.MTW ***
C1 C2 C3 C4 C5 C6 C7 C8 C19 C20
2-Day 4-Day 14-Day
1 270 218 156
2 236 234 *
3 210 214 242
4 142 116 *
5 280 200 *
6 272 276 256
7 160 146 142
8 220 182 216
9 226 238 248
10 242 288 *
11 186 190 168
12 266 236 236
13 206 244 *
14 318 258 200
15 294 240 264
16 282 294 *
17 234 220 264
Project Man...
Current Worksheet: Choleste.MTW
16:23
Start
Inbox - Microsoft Outlook
MINITAB - Untitled
ManualdeMinitab_Julio.do...
MINITAB Help
16:23

Como se observa en el resultado, en la pantalla de Session se reflejará la media muestral, la desviación muestral, el intervalo de confianza al 95% (puede modificarse desde **'Options'**) y el p-valor que permitirá resolver el test (por debajo de 0,025 en el caso bilateral, dirá que es muy poco probable que la media es realmente la especificada inicialmente).

En la práctica, resulta muy frecuente no conocer ni la media ni la desviación típica de la variable, dato este último necesario para realizar un test para el valor medio de la variable. En estos casos, la manera de resolver este contraste es a través de:

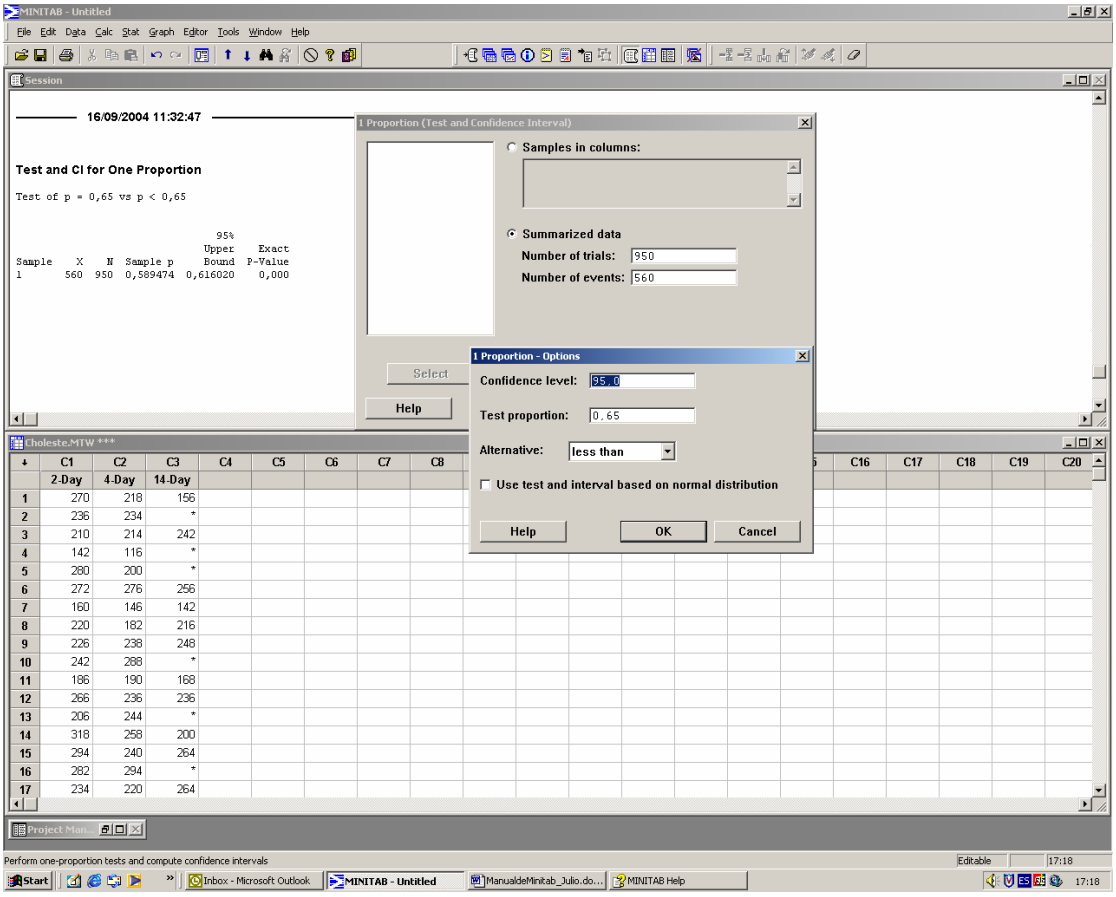
5.2.2 Stat/Basic Statistics/1-sample t

Este situación se resuelve de manera idéntica a la anterior, excepto que no se pregunta por la desviación típica de la variable considera y, por tanto, el test se resuelve utilizando la desviación muestral.

5.2.3 Stat/Basic Statistic/1 Proportion

También es muy frecuente en la práctica plantearse cuestiones sobre la proporción de elementos de la población que cumplen determinada condición (proporción de artículos defectuosos, proporción de alumnos aprobados, proporción de clientes satisfechos,...) En este caso, interesará resolver un test para comprobar si realmente la proporción es la que se supone y/o intervalo de confianza para tal proporción. Al igual que en el caso del test para la media, la ventana de diálogo preguntará el nombre de la variable donde se encuentran los datos y el valor de p para realizar el contraste. Por defecto, lo resolverá de manera bilateral y calculará el intervalo de confianza, también bilateral al 95%. Todo esto puede modificarse desde **'Options'**.

La referencia en la ventana de diálogo a que resuelva basándose en una distribución normal es debido a que MINITAB resuelve el test y el intervalo de manera exacta y no a través de la aproximación de la binomial a la normal, como normalmente se resuelve cuando los cálculos se realizan a mano.



5.2.4 Stat/Basic Statistics/2-Sample t

En esta prueba se trata de comprobar la hipótesis de la no existencia de diferencias significativas entre las medias de dos muestras distintas:

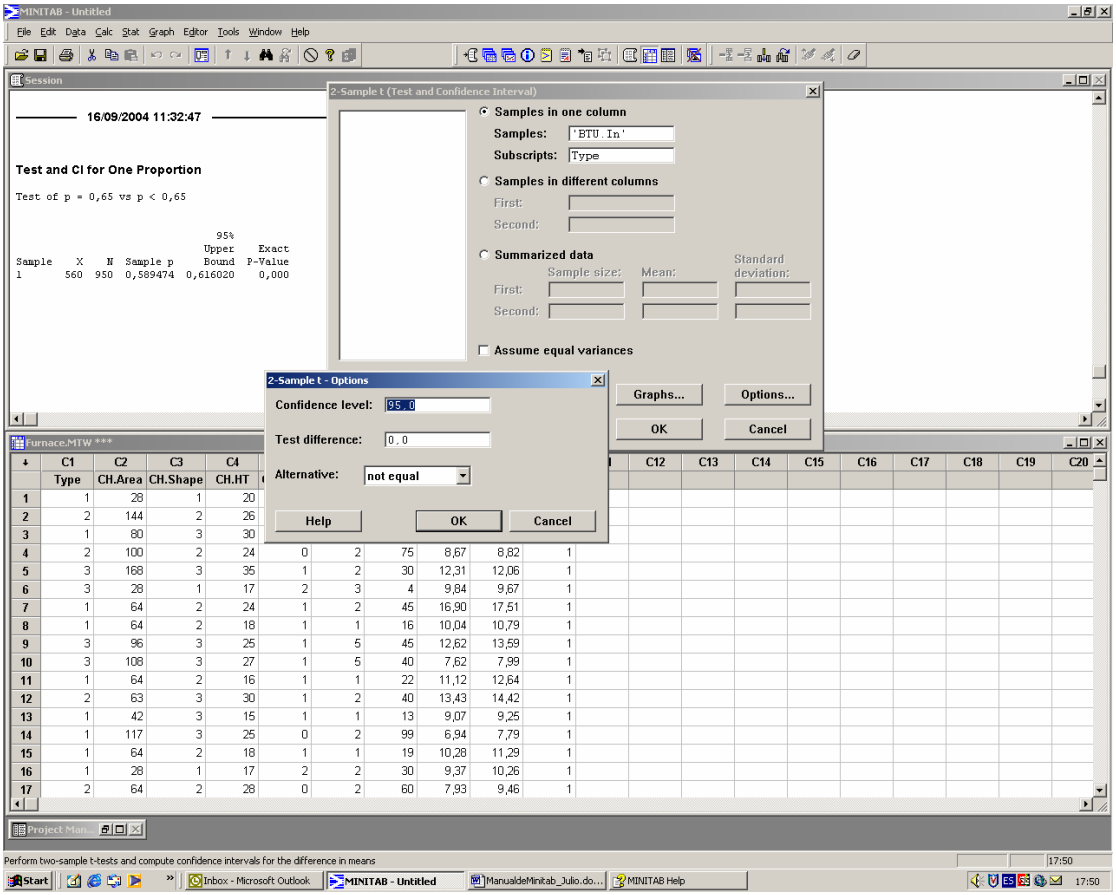
$$H_0: \mu_0 = \mu_1$$
$$H_1: \mu_0 \neq \mu_1$$

Es decir, ahora se dispone de dos muestras procedentes de dos poblaciones distintas, supuestas distribuidas normalmente e independientes y se trata de comprobar si existen o no diferencias significativas entre ambas.

El cuadro de diálogo inicial permite tener los datos en la hoja de diversas maneras: las dos muestras en la misma columna y otra columna que identifique cada muestra (samples in one column); o, cada muestra en una columna (Samples in diffrente columns); o los datos resumidos (solo se dispone de medias y desviaciones muestrales). Una vez elegida la forma adecuada a la disposición de los datos, se indicar si se quiere asumir la igualdad de varianzas entre variables (en caso de que se

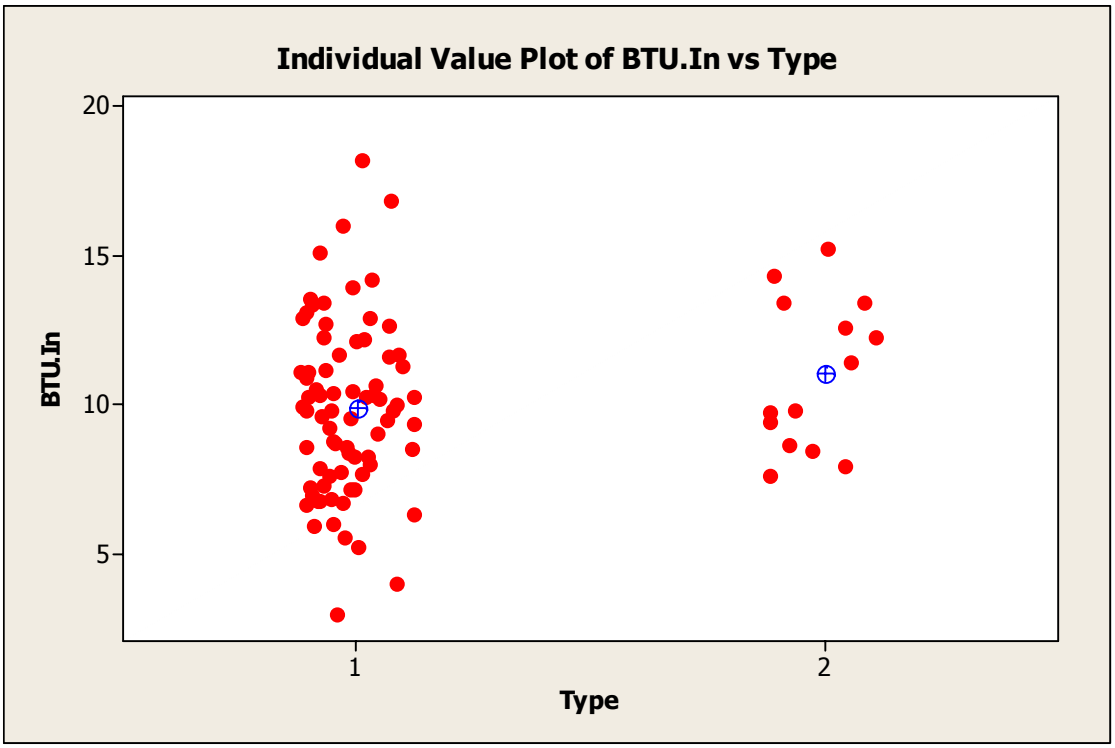
sepa). Como en los test anteriores se podrá representar los datos muestrales gráficamente (eligiéndolos **en 'Graphs'**) y, por defecto, el intervalo d3e confianza para la diferencia de medias se construirá con un nivel de confianza del 95% y la hipótesis alternativa del test será bilateral. Todo esto se puede modificar desde '**Options**'.

Veamos un ejemplo resuelto de comparar el consumo energético de dos calentadores a gas:



Two-Sample T-Test and CI: BTU.In; Type

```
Two-sample T for BTU.In

Type    N    Mean   StDev  SE Mean
1      76    9,85    2,90     0,33
2      14   11,04    2,53     0,68


Difference = mu (1) - mu (2)
Estimate for difference:  -1,18267
95% CI for difference:  (-2,76237; 0,39704)
T-Test of difference = 0 (vs not =): T-Value = -1,57  P-Value = 0,134  DF =
19
```



5.2.5 Stat/Basic Statistics/Paired t

Al igual que en el apartado anterior se trata de contrastar la hipótesis nula de la no-existencia de diferencias significativas entre las medias de dos muestras, pero en este caso con datos apareados, es decir, con los mismo sujetos en dos situaciones diferentes, o con sujetos distintos en ambos grupos, pero que sean comparables par a par respecto de alguna característica de experimentación.

Por ejemplo, imagine que tiene dos grupos de empresas que se diferencian por su actividad: variable SECTOR (sector servicios y sector industria). Si se quiere contrastar la hipótesis de si los beneficios promedio en el año 1997 difieren significativamente entre empresas de un sector y otro, se realizaría un test de hipótesis con dos muestras independientes. Si lo que se quiere en cambio, es contrastar la hipótesis nula de la no existencia de diferencias significativas entre los beneficios obtenidos en el año 1996 y los obtenidos en el año 1997, se trataría de un test de hipótesis con datos apareados, en este caso con los mismos sujetos (empresas) en distintas situaciones (beneficios en cada año).

Por último, si se quiere comparar los beneficios obtenidos por estas empresas con otras homólogas, pero alemanas por ejemplo, se tendría un test de hipótesis con datos apareados con sujetos distintos pero comparables.

Para resolver este caso, los datos deben haberse guardado en columnas diferentes o tener calculadas las diferencias, dos a dos, y su diferencia media y su desviación típica disponiendo de esta información agrupada. El resto, al igual que la interpretación de los resultados, se efectúa de manera idéntica a las anteriores.

5.2.6 Stat/Basic Statistics/ 2 Proportions

Lo mismo que los casos anteriores pero para poder comparar las proporciones de elementos con cierta característica en dos poblaciones. Al igual que en casos anteriores se dan opciones de tener introducidos los datos, e incluso se permite resolver sin introducirlos, únicamente teniendo contabilizados los éxitos y los tamaños de cada muestra.

5.2.7 Stat/Basic Statistics/2 Variances

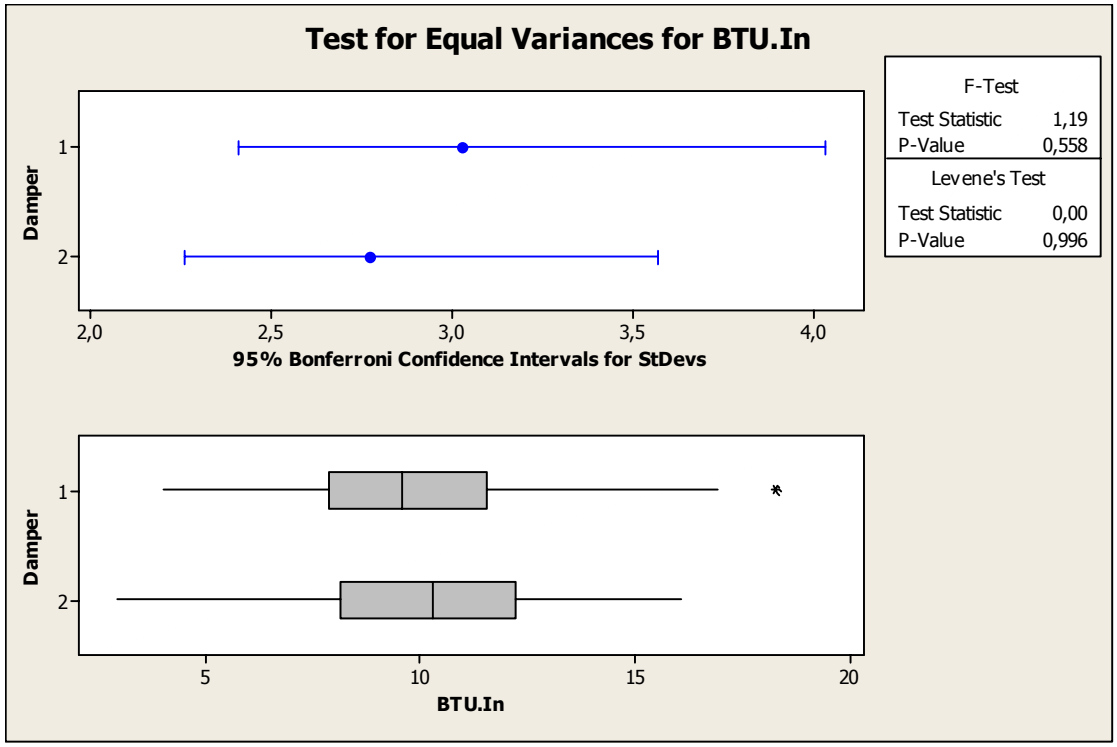
Este test es el que se realiza para comprobar si dos variables tienen la misma varianza, hipótesis de partida bastante frecuente en muchas técnicas estadísticas. Como en los test anteriores, primeramente hay que decir como se tienen guardado los datos (en la misma columna, en dos columnas distintas o los datos resumidos). Además en **'Options'** se puede cambiar el nivel de confianza del intervalo que calcula y en **'Storage'** se puede pedir que almacene los valores de las varianzas calculadas y de los intervalos en las primeras columnas de la hoja de datos.

El test dará como resultado por defecto, el cálculo de un intervalo de confianza para el cociente de varianzas realizado con un nivel de confianza del 95%, y la aplicación de un test para comparar las varianzas de cada muestra (el basado en la distribución F, si las muestras provienen de distribuciones normales) y el de Levene, para el resto de distribuciones continuas. La aceptación de la igualdad de varianzas se realizará, como en otras ocasiones, en función del p-valor calculado.

Al resultado numérico aparecido en la Session, le acompaña un gráfico donde se representan los intervalos de confianza calculados y los gráficos BoxPlot de cada muestra.

Test for Equal Variances: BTU.In versus Damper

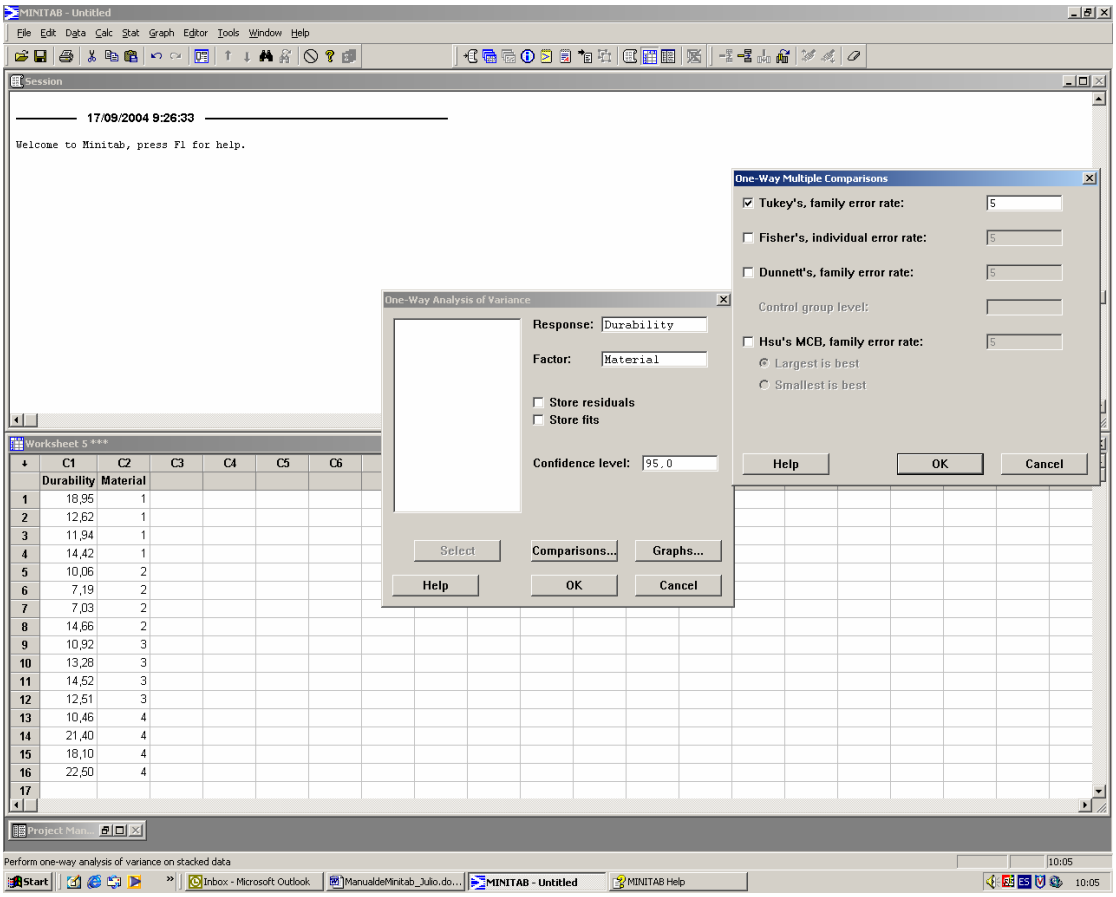
```
95% Bonferroni confidence intervals for standard deviations

Damper    N     Lower     StDev     Upper
     1   40   2,40655   3,01987   4,02726
     2   50   2,25447   2,76702   3,56416

F-Test (normal distribution)
Test statistic = 1,19; p-value = 0,558

Levene's Test (any continuous distribution)
Test statistic = 0,00; p-value = 0,996
```



5.2.8 Comparación de más de dos medias; Análisis de Varianza

Tan frecuente como querer comparar los valores medios de dos muestras (referidas a datos de dos máquinas o dos tratamientos distintos o 2 poblaciones,...) es el de disponer de más de dos muestras para la comparación (3 ó más máquinas, 3 ó más tratamientos distintos, 3 ó más poblaciones,...) Para esta situación y en el caso más sencillo, en el que solo se considere la influencia de un factor en la diferencia, este estudio se debe resolver a través de **Stat/ANOVA/One-Way.**

Imaginemos el siguiente ejemplo. Se desea comprobar la diferencias en la fiabilidad media de unas piezas (en términos de sus meses de vida hasta el primer fallo) de acuerdo al material con el que han sido fabricados. Es decir, la comparación se efectuará teniendo en cuenta que el único factor que puede influir en la respuesta es el material:



Por defecto, el resultado de resolver este estudio será el del p-valor asociado al test correspondiente que nos hará decidir si se observan diferencias significativas en la duración media, de acuerdo al material con el que se ha fabricado la pieza o no (p-valor < 0,05, por defecto, indicaría que las medias son estadísticamente distintas,). En **'Comparisons'** se permite elegir un método (se ofrecen tres distintos) para calcular intervalos de confianza para las diferencias entre medias (intervalos necesarios en el caso de que se observen diferencias significativas) y en **'Graphs'** se pueden representar los residuos para validar el método aplicado para la resolución (paso obligado).

Para este ejemplo, los resultados han sido los siguientes:

One-way ANOVA: Durability versus Material

```
Source     DF      SS     MS      F      P
Material    3   146,4   48,8   3,58  0,047
Error      12   163,5   13,6
Total      15   309,9

S = 3,691   R-Sq = 47,24%   R-Sq(adj) = 34,05%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N    Mean  StDev  ---------+---------+---------+---------+
1       4  14,483  3,157            (-------*-------)
2       4   9,735  3,566  (-------*--------)
3       4  12,808  1,506         (-------*-------)
4       4  18,115  5,435                   (-------*-------)
                          ---------+---------+---------+---------+
                                10,0      15,0      20,0      25,0

Pooled StDev = 3,691


Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Material

Individual confidence level = 98,83%


Material = 1 subtracted from:

Material    Lower  Center   Upper  ------+---------+---------+---------+---
2         -12,498  -4,748   3,003      (------*-------)
3          -9,426  -1,675   6,076         (------*-------)
4          -4,118   3,632  11,383              (-------*------)
                             ------+---------+---------+---------+---
                                 -10         0        10        20


Material = 2 subtracted from:

Material   Lower  Center   Upper  ------+---------+---------+---------+---
3         -4,678   3,073  10,823            (-------*-------)
4          0,629   8,380  16,131                  (------*-------)
                            ------+---------+---------+---------+---
                                -10         0        10        20


Material = 3 subtracted from:

Material   Lower  Center   Upper  ------+---------+---------+---------+---
4         -2,443   5,308  13,058               (------*-------)
                            ------+---------+---------+---------+---
                                -10         0        10        20
```

El p-valor del análisis de varianza sale 0,047 < 0,05, luego estadísticamente se observan diferencias significativas en la fiabilidad de las piezas de acuerdo al material con la que se ha fabricado. Como las pruebas se han realizado con 4 materiales distintos, se requiere calcular los intervalos de confianza para diferencias de medias dos a dos, para concluir entre cuáles de estos materiales existe esta diferencia (entre

todos, o solo entre algunos de ellos). Aquellos intervalos que no contengan el valor 0 serán donde se encuentren los niveles con diferencias significativas. En el ejemplo anterior, el único intervalo que no contiene al cero es el que estima la diferencia entre la media del material 2 y la media del material 4. Luego se puede concluir que el uso de uno u otro material influye a la hora de tener en cuenta la durabilidad del producto. Y entre los materiales probados las diferencias se encuentran entre el 2 y el 4. Entre el resto, no se observan diferencias significativas (podría decirse entonces que la variabilidad observada en esos casos puede deberse a la aleatoriedad)

Los intervalos se han calculado de acuerdo al método Tuckey (uno de los más utilizados). Pueden considerarse también el resto de métodos que propone MINITAB (ver la ayuda)

6 TAMAÑOS DE MUESTRAS

Hasta el momento y a lo largo de esta pequeña guía, se han explicado herramientas y técnicas para poder analizar los datos de una muestra procedentes de una población. Es decir, se ha partido del hecho de que los datos están ya tomados. En cambio, en los casos reales, uno debe partir del problema teniendo que tomar los datos para resolverlo y el primer problema real al que uno se enfrenta es cuántos valores observar o medir para empezar a analizar. "¿Cuántos valores hay que observar para poder dar los resultados con cierto rigor?" es, por tanto, una de las preguntas que más frecuentes que se plantea uno cuando va a comenzar un estudio estadístico.

MINITAB ha incorporado en sus últimas versiones una opción que resuelve esta pregunta, que se accede a través de la ruta **Stat/Power and Sample Size**.

6.1 Stat/Power and Simple Size

Accediendo a esta opción, se puede observar que MINITAB da respuesta en la búsqueda del tamaño necesario de muestra diferenciando el análisis posterior que se hará con dicha muestra. Esto es así porque realmente cuando uno se pregunta "¿Cuántos valores hay que observar para poder dar los resultados con cierto rigor?" a continuación también debe preguntarse "¿qué es lo que desea comprobar a partir de los datos que tome?". Es decir, para contestar bien a esta pregunta hay que entender

que si uno toma datos es porque desea comprobar algo: que la media de una respuesta vale tanto, o que las medias de dos tratamientos no pueden considerarse iguales, o que el porcentaje de votantes será menor del esperado,....

En esta guía se explica como utilizar esta herramienta en los casos de que el análisis posterior de los mismos sea uno de los explicados anteriormente:

6.1.1 Stat/Power and Simple Size/1-Sample Z

En la ventana de diálogo que aparece una vez elegida esta opción, aparecen tres casillas, de las cuales hay que rellenar dos y MINITAB calcula la tercera (por ejemplo, en el caso de querer conocer el tamaño de la muestra hay que introducir valores en el valor de potencia y en el de diferencia mínima) Los tres valores son relativos al test de hipótesis para la media de una población cuando su desviación típica es conocida.

Es decir, que si por ejemplo, se desea conocer cual es el tamaño de muestra óptimo para saber si el peso medio de las piezas que fabrica una máquina son de 200 grs., sabiendo que se tiene una desviación típica de 10 grs. es necesario plantearse con antelación qué diferencia mínima se está dispuesto a asumir y con qué probabilidad. Es decir, el tamaño de la muestra dependerá de si se desea detectar una diferencia en la media de 5 unidades o de 15. Por eso, MINITAB obliga a introducir la diferencia que se desea detectar y la probabilidad mínima la que se desea detectar.

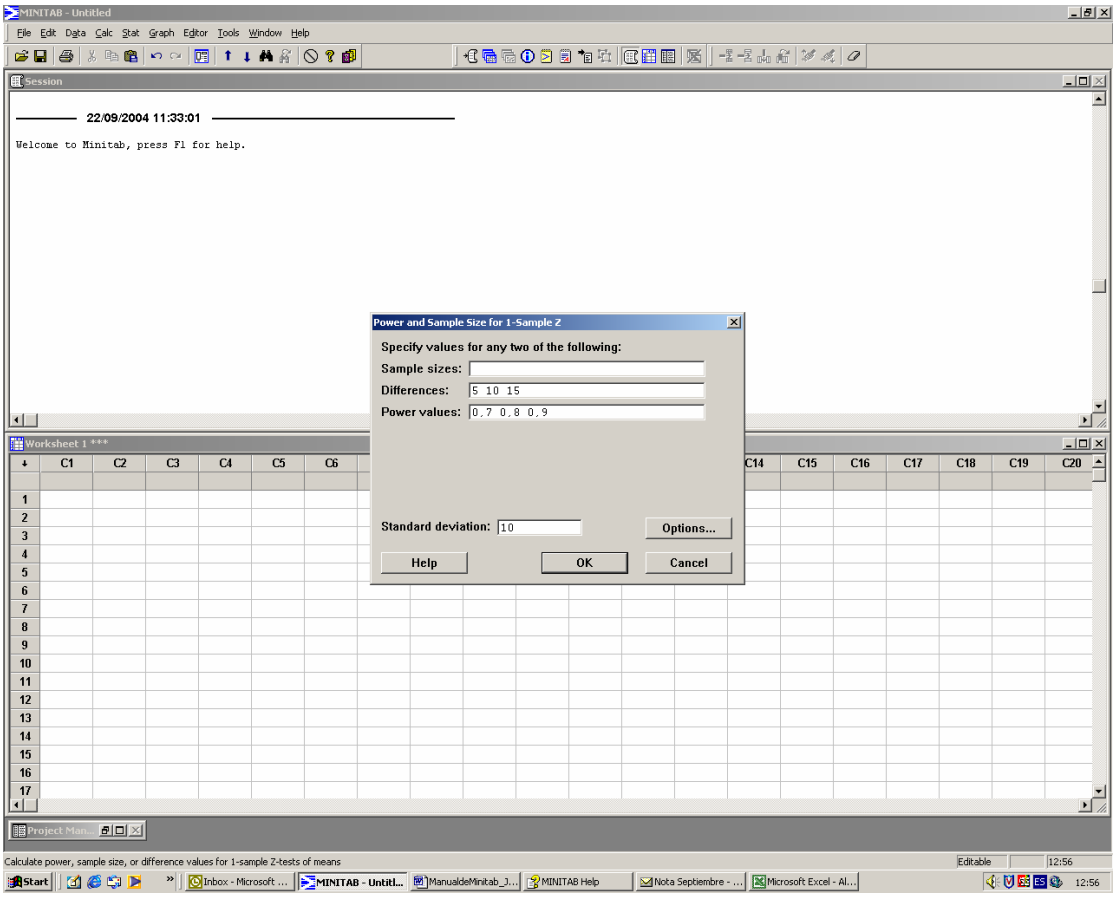
Así, si uno quiere tomar una muestra para comprobar que le peso medio no es de 200 grs., evidenciando diferencias de 5 grs. por lo menos el 80% de las veces, deberá introducir como valor de potencia 0,8 y como valor de diferencia mínima 5:

Power and Sample Size

```
1-Sample Z Test

Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0,05  Assumed standard deviation = 10

            Sample  Target
Difference    Size   Power  Actual Power
         5      32     0,8      0,807430
```

El resultado dice que bastará tomar los pesos de 32 piezas elegidas al azar de la producción de esa máquina para poder detectar con una probabilidad de 0,8 una desviación de la media superior o igual a 5 grs.

Con frecuencia resulta interesante poner varias opciones de diferencias y potencias y decidir en vista a los resultados:

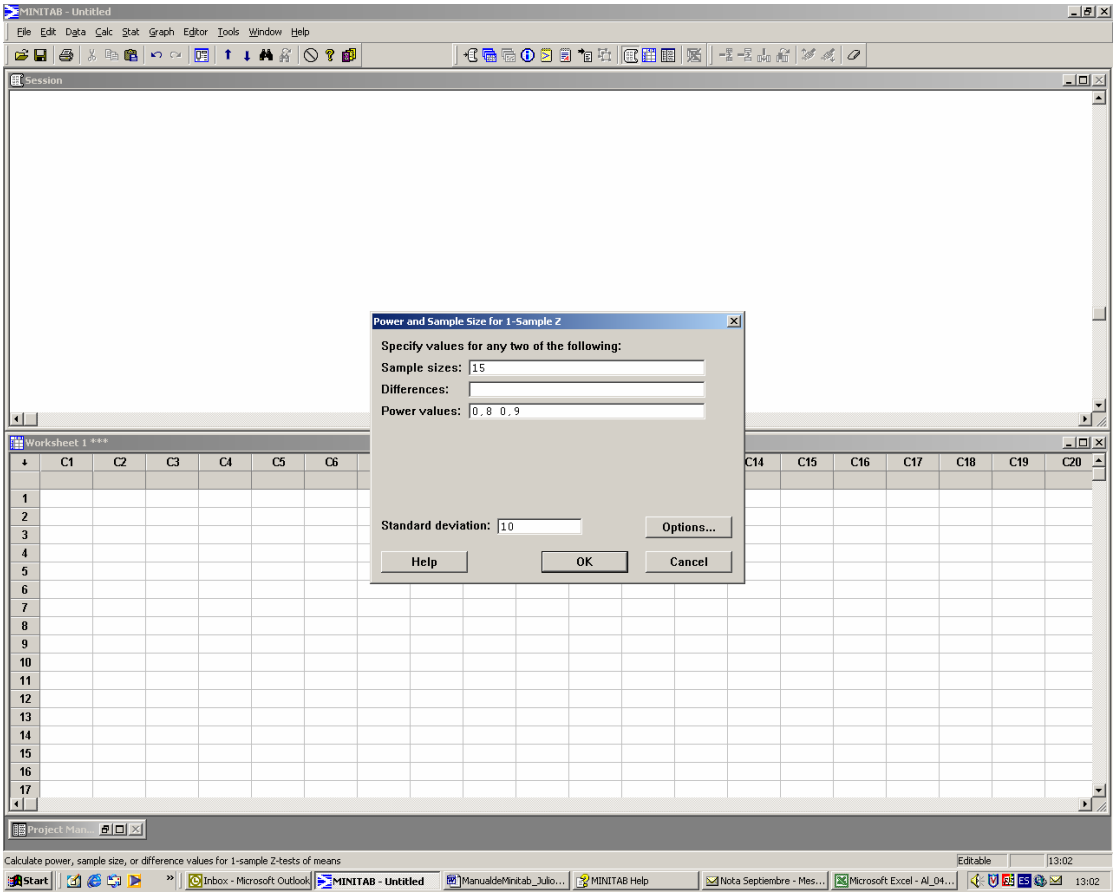


Power and Sample Size

```
1-Sample Z Test

Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0,05  Assumed standard deviation = 10


            Sample  Target
Difference    Size   Power  Actual Power
         5      25     0,7      0,705418
         5      32     0,8      0,807430
         5      43     0,9      0,906375
        10       7     0,7      0,753578
        10       8     0,8      0,807430
        10      11     0,9      0,912556
        15       3     0,7      0,738302
        15       4     0,8      0,850839
        15       5     0,9      0,918362
```

Como se observa en estos resultados, el tamaño de muestra necesario para detectar diferencias superiores a 10 grs. es bastante inferior y en concreto de tomar por ejemplo 7 datos a tomar 8, aumenta un 10% la probabilidad de detectar una mejora de 10 grs. en el peso medio. Si la toma de datos no es excesivamente costosa, seguramente esta sería la mejor elección.

Como he comentado anteriormente, esta misma ventana podrá ser utilizada si, una vez tomada la muestra, uno quisiera conocer qué diferencias son las que detecta con mayor probabilidad. Para esta respuesta debería introducir en el cuadro de diálogo el tamaño de muestra tomado y la probabilidad (power):



Power and Sample Size

```
1-Sample Z Test

Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0,05  Assumed standard deviation = 10


Sample
  Size   Power   Difference
    15     0,8      7,23365
    15     0,9      8,36956
```

Para una muestra de 15 piezas, uno podría detectar muy probablemente (en el 90% de los casos) un desajuste de la máquina en 8,3 grs. respecto de su valor nominal.

6.1.2 Stat/Power and Sample Size/1-Sample T

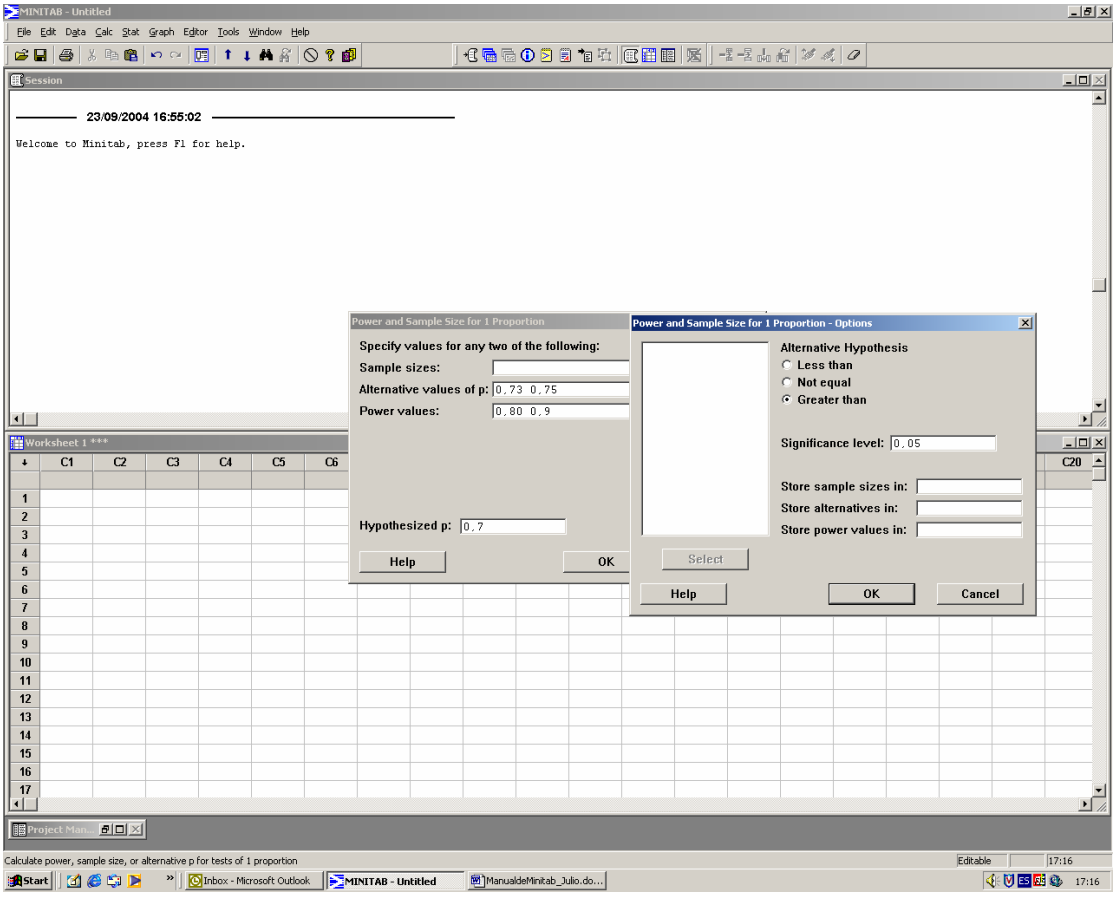
Es exactamente igual que el anterior, solo que no se conoce de entrada la desviación típica de la variable a tratar (lo más común) y, por tanto, el test de hipótesis en el que se basa es en el de comprobar una media con desviación típica desconocida. La ventana de diálogo pide poner el valor de una estimación de esta desviación, que pudiera calcularse, en caso de no conocerla, con alguna muestra de históricos.

6.1.3 Stat/Power and Sample Size/2-Samples T

En este caso, se desea tomar datos para comparar la media de dos poblaciones, por ejemplo, los pesos medios de las piezas que se obtienen de dos máquinas. Vuelven a pedir una estimación de la desviación típica, asumiendo que ambas poblaciones tiene la misma y que pudiera proceder de datos históricos (Pooled StDev)

6.1.4 Stat/Power and Sample Size/1-Proportion

Cuando las hipótesis son sobre la proporción de individuos de la población que verifican cierta propiedad, el tamaño de la muestra se calculará a través de la ruta de 1-proportion, indicando que diferencia mínima quiere detectarse con una alta probabilidad. Por ejemplo, si se quisiera comprobar que la satisfacción de los clientes de la empresa ha aumentado en al menos 3 puntos porcentuales respecto a la del año pasado que fue del 70%, con una probabilidad de al menos 0,80, se tendría que mantener como hipótesis nula que esta proporción es de 0,7 y calcular a cuántos clientes elegidos al azar entre todos los posibles habría que encuestar para poder detectar esta mejora el 80% de los casos:



Power and Sample Size

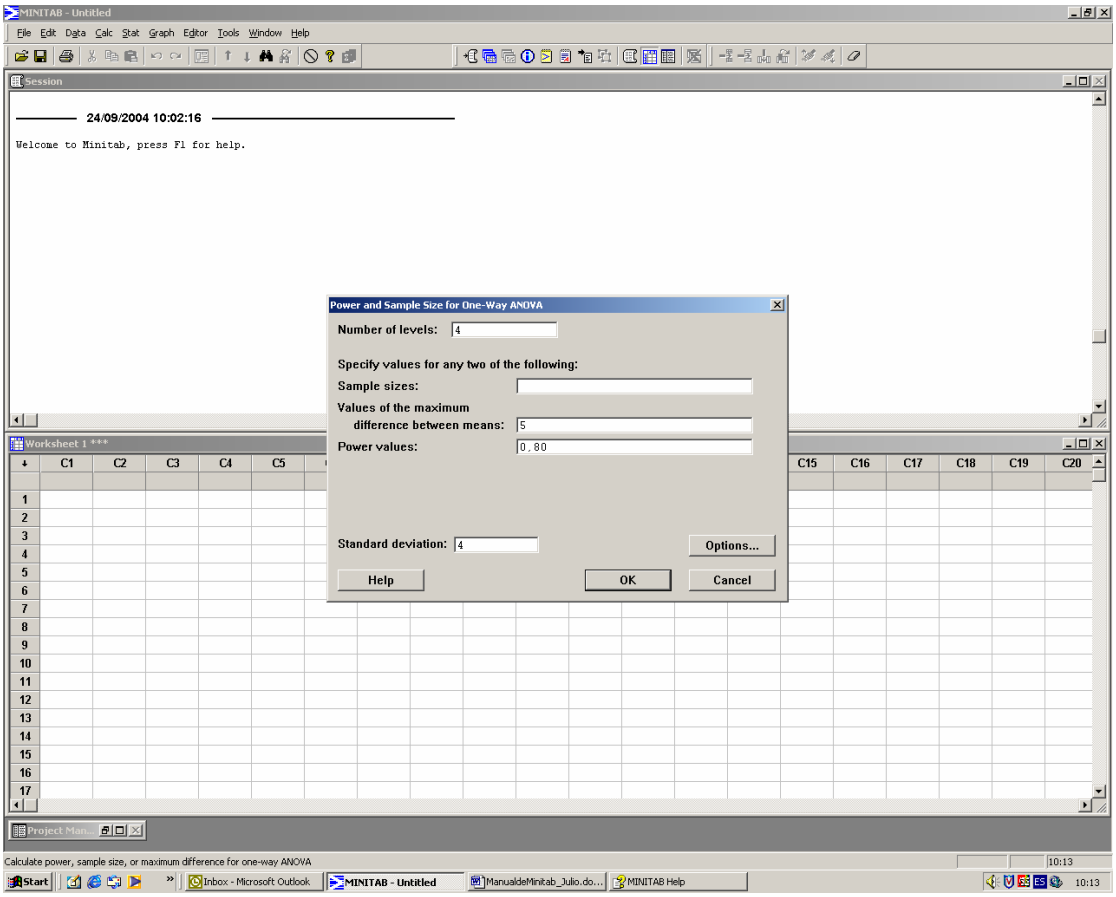
```
Test for One Proportion

Testing proportion = 0,7 (versus > 0,7)
Alpha = 0,05


Alternative  Sample  Target
 Proportion    Size   Power  Actual Power
       0,73    1413     0,8      0,800179
       0,73    1944     0,9      0,900000
       0,75     501     0,8      0,800615
       0,75     686     0,9      0,900358
```

Los resultados indican que se necesitarían 1413 encuestados para darse cuenta, con una probabilidad de 0,80, de que se ha mejorado 3 puntos porcentuales. Con aproximadamente un tercio de encuestas (501), podrías detectar una mejora de un 5%, con la misma probabilidad. Este ejemplo refleja una vez más, lo interesante de esta opción del MINITAB para poder contemplar distintas posibilidades antes de tomar datos.

6.1.5 Stat/Power and Sample Size/2_Proportions

Esta ruta será necesaria cuando se desea comparar las proporciones en dos poblaciones independientes. En este caso, la hipótesis nula que se mantiene es la de que ambas proporciones son iguales y cuando la ventana pide añadir el valor de p_2 (por defecto, 0,5) hay que introducir el valor de la proporción de la hipótesis nula. En la casilla de valores alternativos de p_1 se introducirán las diferencias que se deseen encontrar. Por ejemplo, si un biólogo desea comparar los porcentajes de peces afectados por la polución en dos ríos cercanos para comprobar que uno de ellos ha visto incrementada su proporción en un 5% respecto del otro río, cuando teóricamente ambos tenían un 25% de peces afectados, el valor de p_2 debería ser de 0,25 y el de p_1 de 0,20 y de 0,30. El tamaño de muestra que proporciona este método es igual para cada una de las poblaciones. Lo demás es igual al caso anterior.

6.1.6 Stat/Power and Sample Size/One-way ANOVA



Supongamos que lo que se desea ahora es tomar datos para comparar la producción de 4 máquinas distintas., en relación al peso de las piezas que fabrican. Primeramente habrá tomar aleatoriamente piezas de cada máquina, pesarlas y luego hacer el

ANOVA correspondiente para comparar los pesos medios de las 4 máquinas (ver pág. 51). Luego, lo primero que uno se plantea es cuántas piezas va a inspeccionar de cada máquina. Eligiendo de MINITAB la opción Stat/Power and Sample Size/One-way ANOVA, debe introducirse cuántas medias se desea comparar (Number of levels), la diferencia mínima que desea detectar entre las posibles máquinas diferentes (es decir, debe plantearse qué diferencia mínima desea detectar con el estudio entre la máquina que produzca pesos medios mayores y la que produzca pesos medios menores) y la probabilidad con la que se desea detectar esa diferencia. Asimismo, una vez más, se debe estimar la desviación típica , que debiera tomarse de datos históricos o estudios similares hechos con anterioridad tomando en el caso del ANOVA el valor de 'MS error. Una vez introducidos estos datos, el resultado en la hoja de Session es el siguiente:

Power and Sample Size
One-way ANOVA

Alpha = 0,05 Assumed standard deviation = 4 Number of Levels = 4

SS Means	Sample Size	Target Power	Actual Power	Maximum Difference
12,5	15	0,8	0,800969	5

The sample size is for each level.

7 HERRAMIENTAS DE CALIDAD

Este apartado está en construcción.