Análisis Estadístico. Un enfoque práctico con Statgraphics

Armando Cervantes Sandoval

María José Marques Dos Santos

Patricia Rivera García

Ilustración de portada: Anahy Cruz Mejía

Publicado con apoyo de los proyectos PAPIME EN216403 y EN203503

FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA

Av. Guelatao No. 66, Colonia Ejercito de Oriente,

Delegación Iztapalapa, México, D.F. 09230

ISBN: 970-32-2902-6

Prólogo

Este material es una guía práctica para realizar análisis estadístico de datos, utilizando el software de análisis estadístico Statgraphics. Con ese fin se hace una revisión del "paquete" en un formato de curso, donde se describe paso a paso como utilizar y aplicar algunas de las muchas opciones disponibles, quedando una gran cantidad de ellas todavía por analizar.

Enfocado al uso del software, sin perder formalidad y con el rigor necesario, en cada capítulo se presenta una muy breve explicación estadística, se indica como realizar el análisis correspondiente en Statgraphics y mediante ejemplos se dan algunos criterios que apoyan la interpretación de resultados.

Está dirigido a quienes sabiendo que técnica estadística aplicar requieren de una herramienta confiable que se haga cargo del trabajo de cálculo numérico. No se pretende que sea un libro de estadística, ni siquiera un recetario de estadística y muchos menos un manual de Statgraphics, simplemente se busca que sea una herramienta de apoyo en el quehacer académico de estudiantes, profesores e investigadores. Que los motive e invite a revisar material más formal.

Se recomienda utilizar éste material directamente en el Statgraphics, rehaciendo los ejemplos y ejercicios.

Un reconocimiento especial a los alumnos del curso que dieron origen a estas notas, así como a los de cursos posteriores, por su paciencia y cuidado para realizar algunas correcciones. En especial a las 3 generaciones del Diplomado en Estadística Práctica.

Se agradece a: M en A. Teresa Guerra Dávila, Dr. José Luis Gómez Márquez y Dr. Gerardo Cruz Flores, la minuciosa revisión del este material, así como sus valiosas sugerencias y comentarios.

Finalmente, asumimos la responsabilidad de cualquier error, omisión o mala interpretación que, totalmente sin querer, se presente. Agradeciendo todas las correcciones, comentarios y sugerencias, que ayuden a mejorar este material, al correo electrónico: arpacer@servidor.unam.mx.

Armando Cervantes S.
María José Marques Dos Santos
Patricia Rivera García

Contenido

Pág.

CAPÍTULO 1

BREVE INTRODUCCIÓN A STATGRAPHICS

Como cualquier otro software de Análisis Estadístico el primer paso consiste en entrar y salir de él.

ENTRAR

Seguir la secuencia:
Inicio -> Programas -> STATGRAPHICS Plus -> Sgwin

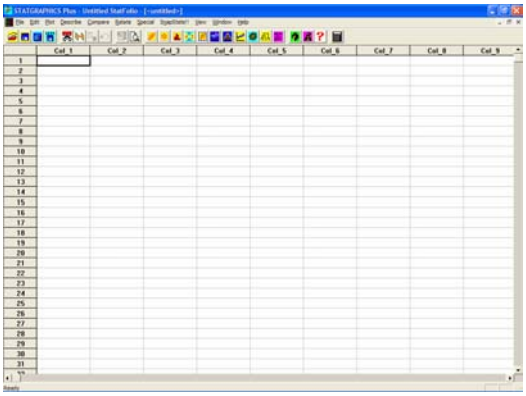


Figura 1. Hoja de Trabajo de STATGRAPHICS.

Salir

Al igual que cualquier aplicación en ambiente Windows, se tienen varias opciones para salir del "paquete"
1. Cerrar con un clic en la caja de la esquina superior derecha, sobre el icono marcado con una fea X.
2. En la caja de diálogo de la esquina superior izquierda de la barra de título, dar un clic y luego seleccionar **cerrar.**
3. Desde la opción de la barra de menú, seleccionar **FILE** y luego **Exit STATGRAPHICS.**

CAPTURA Y EDICIÓN DE DATOS

Este es el paso más importante de cualquier análisis, para el cual Statgraphics cuenta con una hoja de datos, en formato tipo Excel, de manera que sólo basta con "teclear" los datos o importarlos de Excel, con **Copy-Paste**.

Para definir el tipo y tamaño de datos, se recomienda la secuencia:

a) Dar un clic sobre el identificador de columna, por ejemplo **Col_1**.
b) Una vez seleccionada la columna dar un clic derecho y seleccionar del menú flotante que aparece la opción **Modify Column** (figura 2).
c) Una vez que aparece la caja de diálogo de la figura 3, seleccionar las opciones adecuadas, como nombre y tipo de la variable a ingresar en esa columna. También puede llegar hasta aquí dando un doble clic en el paso a.

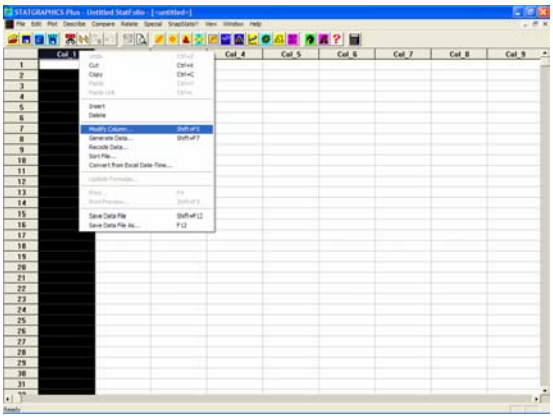


Figura 2. Menú flotante para definir el tipo de variable y asignarle nombre.

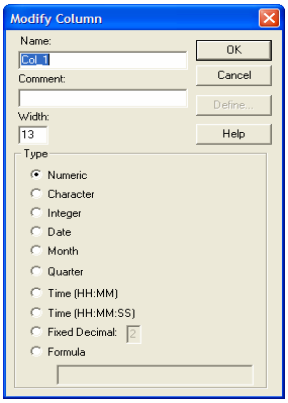


Figura 3. En esta caja de diálogo se puede definir el tipo de dato (se tienen 10 opciones). También se le da un nombre a la variable, se define la amplitud y se le puede agregar un comentario.

NOTA: Los datos se pueden obtener de cualquier software que tenga el formato de hoja electrónica de cálculo y sólo basta con un "copy (cut)" y "paste" para tenerlos en STATGRAPHICS. Si tiene algún problema se recomienda "triangular" con Excel o alguna otra hoja de cálculo.

Al dar clic sobre el identificador de columna, también se puede resaltar la opción **Generate Data** (figuras 4 y 5) para transformar los datos.

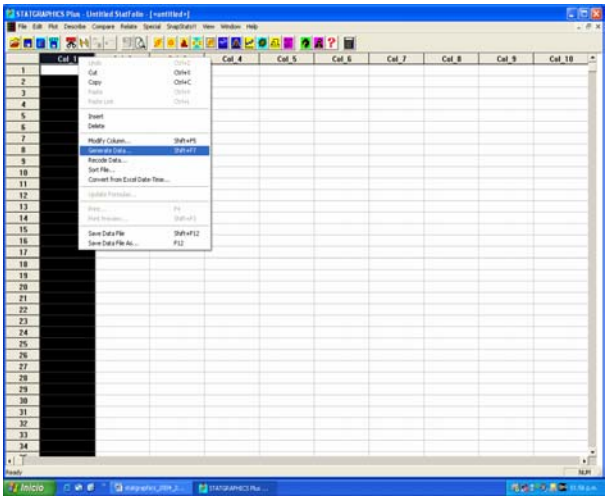


Figura 4. Menú flotante para definir generador de datos.

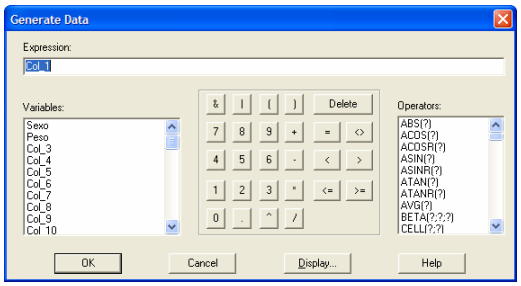


Figura 5. Esta caja de diálogo permite definir nuevas variables a partir de las ya existentes, aplicando las operaciones aritméticas básicas y algunas funciones matemáticas integradas.

Ejemplos para manejo de datos

La distribución de alumnos en las 7 carreras de la Facultad de estudios Superiores Zaragoza es aproximadamente:

Carrera	Número de alumnos
Biología	700
Enfermería	850
Ingeniería Química	400
Medicina	1600
Odontología	1700
Psicología	1800
Q.F.B.	1130

Para generar el archivo se deben considerar dos variables: una llamada **Carrera** de tipo **Character** y otra de nombre **Alumnos** de tipo **Numeric**, entonces se tienen 2 columnas con 7 datos cada una.

Los datos se "teclean" directamente en la hoja de datos y el nombre de las variables y tipo de datos se modifican siguiendo la secuencia.

1. Dar un clic sobre el identificador de la columna, lo que equivale a seleccionar la columna (un doble clic equivale al paso 2).

2. Dar un clic derecho para que aparezca un menú flotante.

3. Seleccionar la opción **Modify Column** y utilizar la caja de diálogo que aparece (figura 3), para definir las mejores opciones de trabajo.

Como segundo ejemplo se tiene:

Peso en Kg de una muestra al azar de 50 estudiantes de una universidad, tomados de sus registros médicos.

Género	M	M	M	F	F	M	F	M	M	F
Peso	89	93	96	64	68	98	69	102	95	60

Género	F	M	F	M	M	F	F	M	M	F
Peso	49	75	54	79	81	56	59	84	85	60

Género	M	M	F	M	F	F	M	M	M	M
Peso	94	88	59	84	58	58	81	79	77	74

Género	F	M	M	F	M	F	M	F	M	M
Peso	51	72	73	54	78	55	77	58	81	83

Género	F	M	F	M	M	F	F	M	M	M
Peso	71	92	60	85	98	66	65	108	88	92

Para generar el archivo se deben considerar dos variables: una llamada **Género** de tipo **Character** y otra de nombre **Peso** de tipo **Numeric**, entonces se tienen 2 columnas con 50 datos cada una.

Al igual que en caso anterior, los datos se "teclean" directamente en la hoja de datos y el nombre de las variables y tipo de datos se modifican siguiendo la secuencia.

1. Dar un clic sobre el identificador de la columna, lo que equivale a seleccionar la columna (un doble clic equivale al paso 2).

2. Dar un clic derecho para que aparezca un menú flotante.

3. Seleccionar la opción **Modify Column** y utilizar la caja de diálogo que aparece (figura 3), para definir las mejores opciones de trabajo.

Salvar o Guardar un archivo de datos

El siguiente paso consiste en almacenar o "guardar" los datos en un archivo, lo que se logra con la secuencia.

File -> Save As -> Save Data File As

Enseguida aparece una caja de diálogo donde se indica el nombre del archivo.

Se debe aclarar que el archivo se guarda con un formato STATGRAPHICS, por lo que sólo podrá "abrirse" en este "paquete". Aunque también puede guardarse como archivo texto, TXT, o XML.

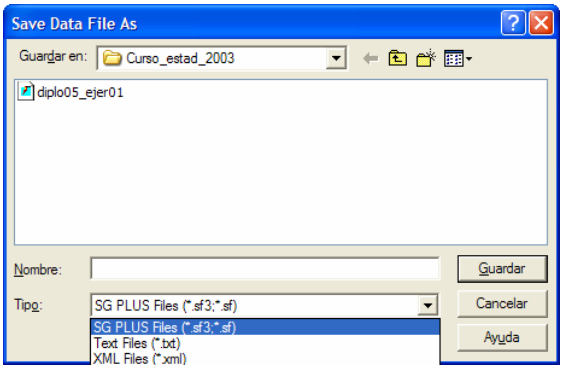


Figura 6. Caja de diálogo para asignarle nombre a un archivo de datos.

NOTA: STATGRAPHICS funciona con una metáfora de escritorio de trabajo, donde cada tipo de ventana es parte de este escritorio, de manera que para guardar todas las ventanas se puede utilizar **SAVE STATFOLIO AS**.

Si se realizan modificaciones posteriores se puede utilizar la opción SAVE, en lugar de SAVE AS.

Procesando Datos

El primer paso de un análisis consiste en explorar los datos de manera gráfica, para esto Statgraphics cuenta con varias opciones. Por ejemplo, de la barra de menú se pueden seguir las secuencias.

a) Plot -> Business Chart -> Pie Chart

b) Plot -> Business Chart -> Bar Chart

Las dos gráficas anteriores son para datos de tipo categórico (nominales u ordinales), como es el caso del ejemplo de las carreras de la FES Zaragoza.

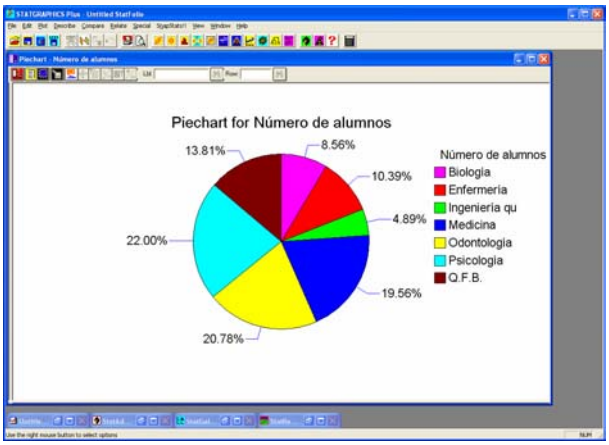


Figura 7. Diagrama de Pie (pastel) para el número de alumnos de las carreras de la FES Zaragoza.

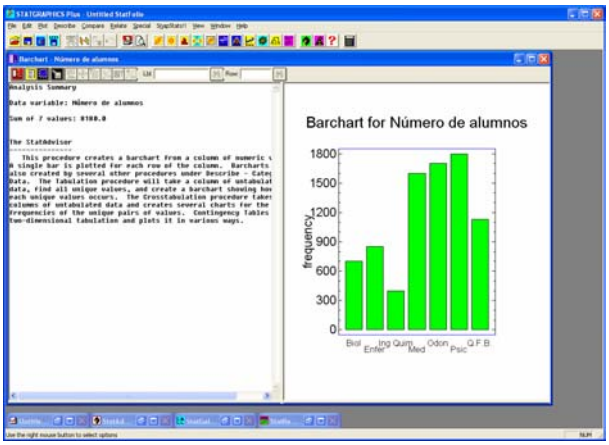


Figura 8. Diagrama de Barras para el número de alumnos de las carreras de la FES Zaragoza.

Para datos cuantitativos se puede seguir la secuencia

c) Plot -> Exploratory Plots -> Box and Whisker Plot

d) Plot -> Exploratory Plots -> Multiple Box and Whisker Plot

e) Plot -> Exploratory Plots -> Frequency Histogram

Estas dos gráficas son para datos de tipo continuo o de escala.

Esta última opción presenta una caja de diálogo que permite seleccionar las variables a trabajar (figura 9).

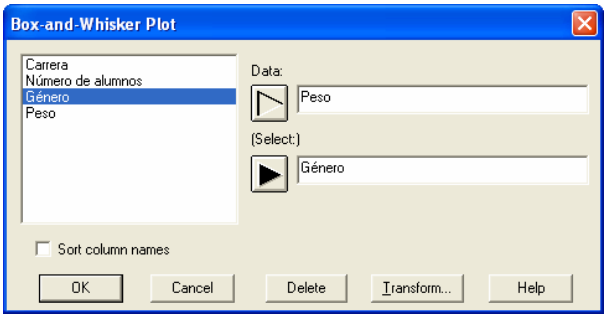


Figura 9. Selección de variables para un Diagrama de caja y bigote (también conocida, como caja y alambre).

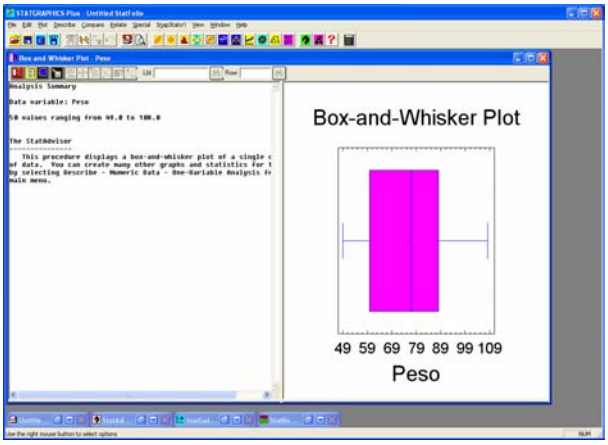


Figura 10. Diagrama de caja y bigote.

Al dar clic en el botón derecho sobre la gráfica se puede entrar a **pane options** para cambiar las opciones de vertical a horizontal o de diagrama con o sin muesca.

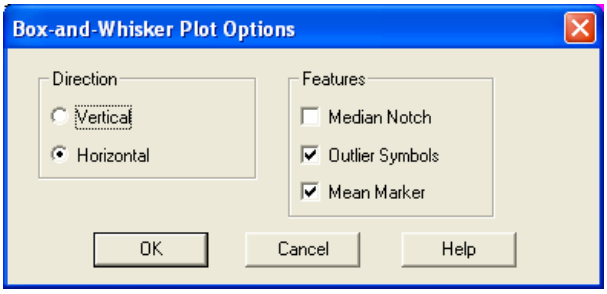


Figura 11. Pane Options para Diagrama de caja y bigote.

Para realizar un diagrama de cajas múltiple, para ambos géneros, se selecciona

Plot -> Exploratory Plots -> Multiple Box and Whisker Plot

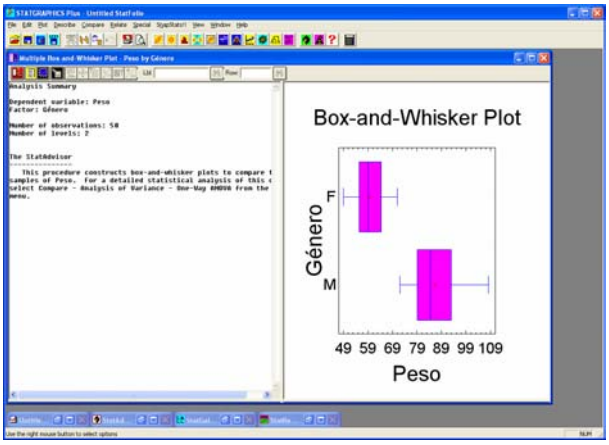
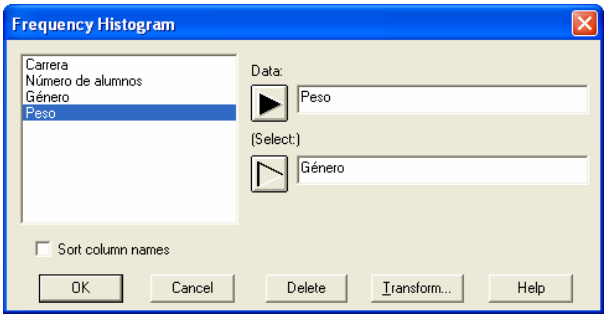


Figura 12. Diagrama de caja y bigote múltiple.

Al igual que el anterior se tienen las opciones con y sin muesca y vertical u horizontal

Para realizar el histograma de frecuencias se sigue la secuencia

Plot -> Exploratory Plots -> Frequency Histogram



Al aceptar aparece la ventana de resultados, que se divide en dos partes, una para resultados tabulares y otra para los gráficos (figura 13).

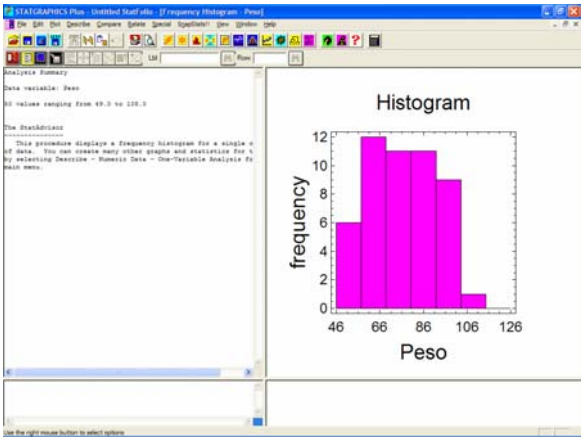


Figura 13. Ventana de salida o resultados.

Análisis de la ventana de resultados

1. Al dar un doble clic sobre las ventanas estas se pueden maximizar o minimizar.

2. Hay una barra de iconos de la ventana de resultados, los más importantes son: un botón rojo que permite abrir el diálogo para seleccionar variables a utilizar en el análisis; un botón amarillo que permite seleccionar las opciones tabulares a trabajar (figura 14), un botón de fondo negro con un gráfico en azul que permite seleccionarlas opciones gráficas del análisis (figura 15). Un botón que muestra un disco flexible, para guardar o almacenar los resultados en disco.

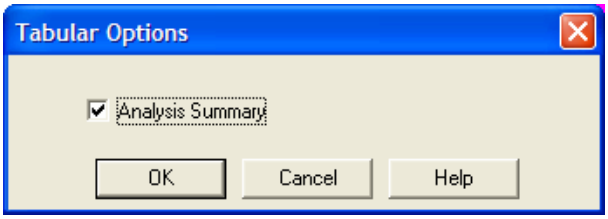


Figura 14. Opciones Tabulares, Las opciones dependen del tipo de análisis que se esté realizando.

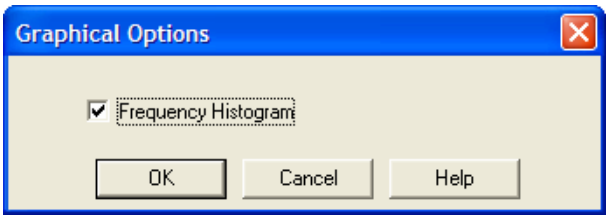


Figura 15. Opciones Gráficas.

STATGRAPHICS tiene la característica distintiva de contar con un "Sistema Experto" que guía el análisis y apoya la interpretación de resultados.

Cada ventana tiene varias posibilidades de trabajo, las cuales se despliegan en un menú flotante que aparece al dar un clic derecho sobre la ventana.

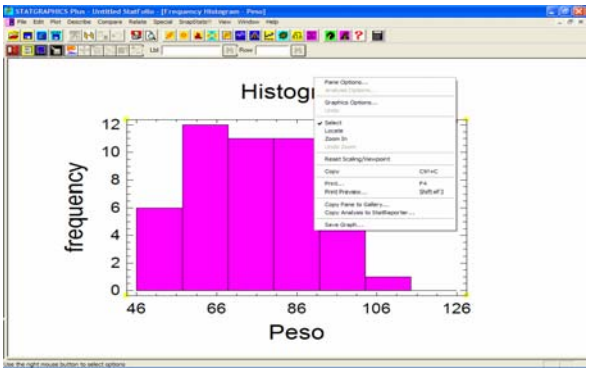


Figura 16. Menú flotante de resultados.

Es importante resaltar las opciones: **Pane Options, Graphics Options y Copy Análisis to StatReporter**.

Pane Options, despliega las opciones de la técnica seleccionada.

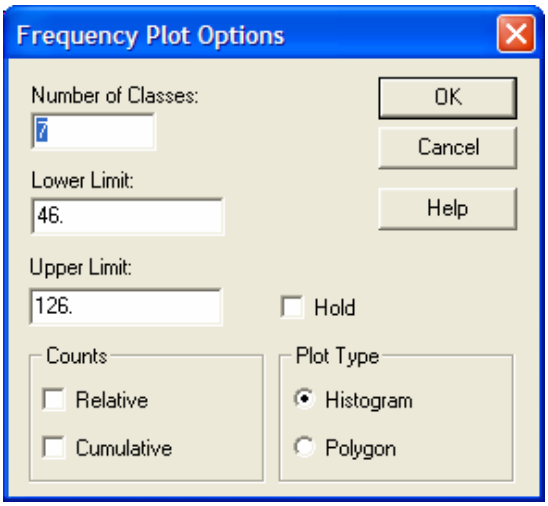


Figura 17. Opciones del Histograma de Frecuencias.

Graphics Options, permite modificar el gráfico, ya sea el título principal, títulos de ejes, tipo y tamaño de fuentes, gráfico en dos o tres dimensiones, así como la orientación de texto.

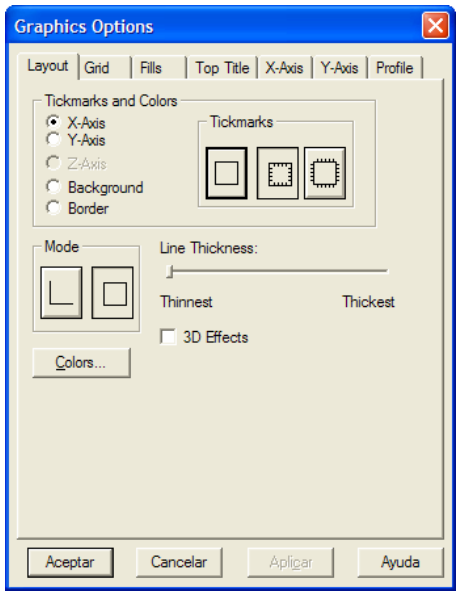


Figura 18. Opciones Gráficas, permite editar un gráfico antes de guardarlo.

Copy Análisis to StatReporter, es la opción que permite almacenar los resultados en un generador de reportes de STATGRAPHICS, equivalente a un editor de texto donde se puede agregar o editar texto.

Almacenar, Editar y Guardar en Disco los Resultados

Como se acaba de mencionar, los resultados se guardan o almacenan al seleccionar **Copy Análisis to StatReporter** (NOTA: Sólo se almacena en la memoria de trabajo de STATGRAPHICS), la información almacenada se puede consultar y modificar al seleccionar de la barra de menú la secuencia:

Window -> StatReporter

Desplegándose un ambiente de edición donde se puede agregar, borrar o modificar la información.

Una vez realizadas todas las modificaciones, la información se almacena en disco al dar un clic derecho y del menú flotante que aparece seleccionar **Save StatReporter As**, entonces aparece un diálogo para darle un nombre al archivo. En este menú también aparece la opción de limpiar el área de StatReporter.

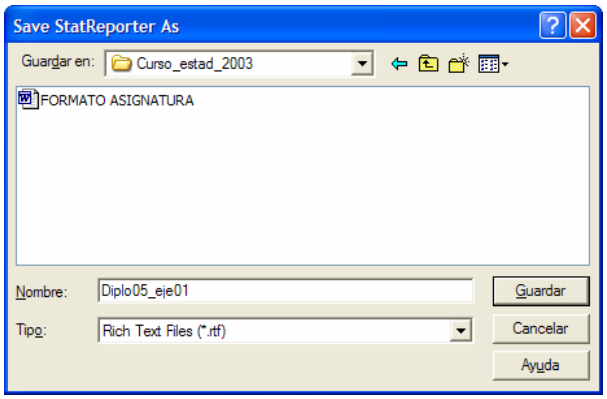


Figura 19 Diálogo para salvar el contenido del StatReporter.

NOTA: El formato de este archivo es RTF (Riched Text Format), el cual se puede "abrir" directamente en cualquier procesador de palabras, como Word de Microsoft.

Es importante practicar este manejo de datos y archivos, ya que son los pasos esenciales para:

1. Ingresar los datos al software de análisis estadístico.

2. Tener datos para seleccionar un análisis de la amplia gama de posibilidades que se encuentra en este "paquete".

3. Almacenar los resultados en un formato que permita realizar un reporte escrito de la manera más rápida y sencilla.

4. Al salvar el archivo StatReporter no se guardan los datos sino solamente los resultados, por ello es necesario salvar los datos como un archivo Datafile.

CAPÍTULO 2

GRÁFICOS CIRCULARES Y DE BARRAS

Estos gráficos se utilizan para trabajar con datos de tipo cualitativo, cuya respuesta se mide mediante conteos o datos de frecuencias. Su manejo se muestra mediante ejemplos.

GRÁFICOS DE PIE, CIRCULARES O DE PASTEL

Ejemplo 1

Se le pidió a los alumnos de primer año del Colegio de Administración de una Universidad que indicaran su campo de estudio preferido, los resultados se presentan en el siguiente cuadro.

Campo	Alumnos
Administración	55
Contabilidad	51
Finanzas	28
Mercadotecnia	82

Los pasos a seguir son:

1. Generar un archivo con 2 columnas, una llamada **campo** de **tipo character** y una llamada **frec_alumnos** de **tipo numérico**. Se tienen sólo 4 filas de datos y se recomienda guardar el archivo de datos en disco, antes de continuar.
2. Seguir, del menú, la secuencia:

Plot ->Business Charts -> Piechart

3. Colocar las variables adecuadas en la caja de diálogo que aparece.

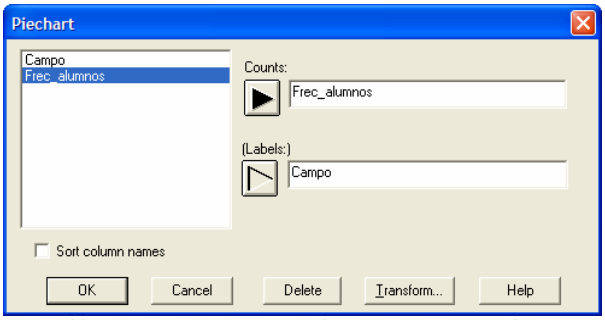
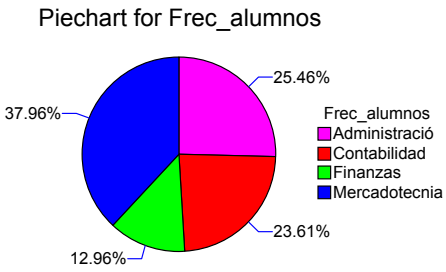


Figura 20 Diálogo para realizar gráficos de pastel, en **Counts** se coloca la variable que tiene los conteos o frecuencias y en **Labels** la variable que identifica las categorías.

4. Presionar el botón **OK**, para obtener los siguientes resultados

Piechart - Frec_alumnos

```
Analysis Summary

Data variable: Frec_alumnos

Sum of 4 values: 216.0


The StatAdvisor
---------------
   This procedure creates a piechart from a
column of numeric values.  Piecharts are also
created under Describe - Categorical Data
Tabulation.  The Tabulation procedure will take
a column of untabulated data, find all unique
values, and create a piechart showing how often
each unique value occurs.
```



Estos son todos los resultados posibles, ya que las opciones tabulares y las opciones gráficas no muestran más resultados.

Para trabajar con el **Pane Options** del gráfico:

1. Dar un doble clic en la ventana de gráficos para maximizarla.

2. Dar un clic derecho y del menú que aparece seleccionar **Pane Options**. De la caja de diálogo que aparece (figura 21 seleccionar las mejores opciones de trabajo.

Se puede definir el tipo de leyenda (**Legends**) que se quiere desplegar en el gráfico.

En cada rebanada también se tienen opciones para el despliegue de etiquetas (**Labels**).

El tamaño del gráfico (**Diameter**), si alguna "rebanada" va separada del resto del gráfico (**Offset Slice #**) y si se presentas líneas que unan las etiquetas con la "rebanada" correspondiente (**Lines**).

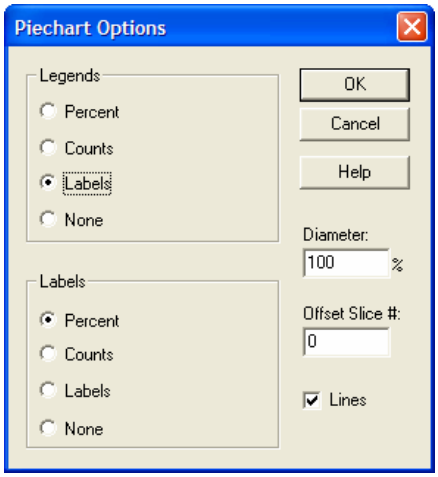


Figura 21. Diálogo Pane Options para gráficos circulares.

Tambien se pueden probar las **Graphics Options**, siguiendo el mismo procedimiento de Pane Options. Con esto aparece un menú en forma de carpetas o folders, donde se pueden modificar los atributos del gráfico (figura 22), como título, tipo y tamaño de letra, así como la orientación del texto.

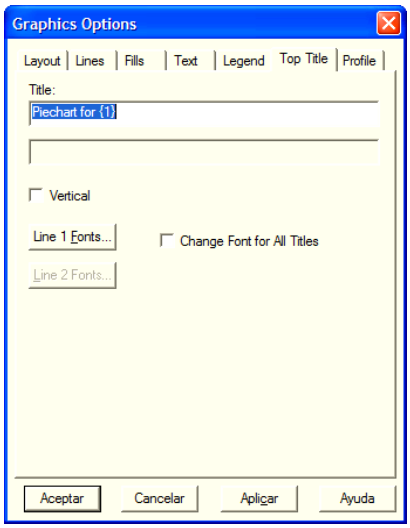


Figura 22. Diálogo Graphics Options para gráficos circulares.

Es importante recordar que los resultados se almacenan en el StatReporter, al dar un **clic derecho** sobre cualquier ventana de resultados y seleccionando **Copy Analysis to StatReporter**.

Estos resultados se guardan en disco en un formato accesible a cualquier procesador de texto al seguir la secuencia**: File -> Save As -> Save StatReporter As**

Ejemplo 2.

Los siguientes datos representan ventas, en millones de dólares, de 20 empresas de la industria de servicios de cuidado de salud (Business Week, 17 de noviembre de 1977).

Empresa	Ventas
Beverly Ent.	805
Coventry	307
Express Scripts	320
HealthSouth	748
Horizon	445
Humana	1968
Int. Health	472
Lab. Corp. Am.	377
Manor Care	274
Medpartners	1614
Novacare	357
Phycor	284
Quest Diag.	374
Quorum Health	393
Sun Healthcare	486
Tenet Haelthcare	2331
U. Wisconsin	389
Univ. Health	362
Vencor	845
Wellpoint	1512

Utilice un gráfico de Pastel para describir este conjunto de datos.

Los pasos a seguir son:
1. Generar un archivo con 2 columnas, una llamada **Empresa** de **tipo character** y una llamada **Ventas** de **tipo numérico**. Se tienen 20 filas de datos y se recomienda guardar el archivo de datos en disco, antes de continuar.

2. Seguir, del menú, la secuencia:

Plot ->Business Charts -> Piechart

3. Colocar las variables adecuadas en la caja de diálogo que aparece.

4. Una vez que aparecen los resultados se puede "jugar" con las opciones del **Pane Options** o de **Graphics Options**.

5. Guardar los resultados en el StatReporter.

6. Almacenar en disco el contenido del StatReporter.

Resultados

Piechart - Ventas

```
Analysis Summary

Data variable: Ventas

Sum of 20 values: 14663.0


The StatAdvisor
---------------
   This procedure creates a piechart from a
column of numeric values. Piecharts are also
created under Describe - Categorical Data -
Tabulation.  The Tabulation procedure will take
a column of untabulated data, find all unique
values, and create a piechart showing how often
each unique value occurs.
```

Interpretación

Este tipo de gráficos sólo permite visualizar quien tiene mayor frecuencia o mayor porcentaje, realizando una comparación relativa entre todas las categorías que se tienen en los datos

GRÁFICOS DE BARRAS

Para los datos del ejemplo 1, de este capítulo, como ya se tienen los datos almacenados basta con “abrir” el archivo, con la secuencia

File -> Open -> Open Data File

Seleccionar el archivo adecuado, definiendo la trayectoria en la que esta almacenado (unidad de disco, carpeta y subcarpeta y nombre del archivo) para llenar la caja de diálogo que aparece (figura 23).

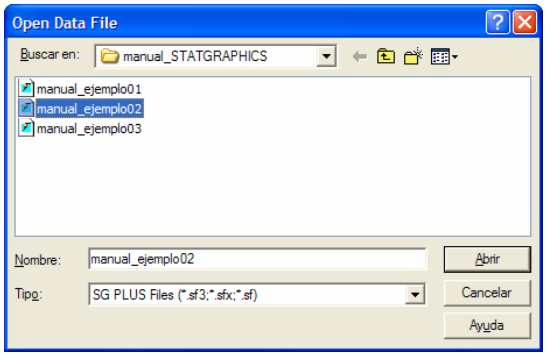


Figura 23. Diálogo para abrir archivos de datos

Una vez que ya se tienen los datos, seguir la secuencia

Plot -> Business Charts -> Barchart

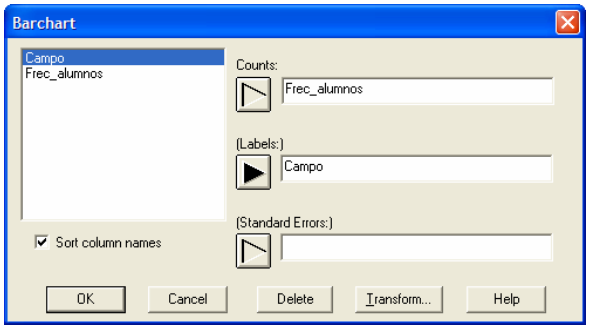
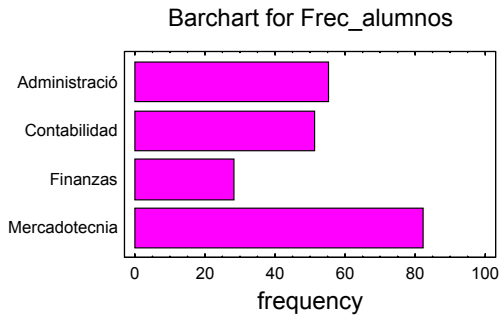


Figura 24. Diálogo para definir variables en un gráfico de barras.

Resultados

Barchart - Frec_alumnos

```
Analysis Summary

Data variable: Frec_alumnos
Sum of 4 values: 216.0
The StatAdvisor
---------------
   This procedure creates a barchart from a
column of numeric values. A single bar is
plotted for each row of the column.  Barcharts
are also created by several other procedures
under Describe - Categorical Data.  The
Tabulation procedure will take a column of
untabulated data, find all unique values, and
create a barchart showing how often each unique
values occurs.  The Crosstabulation procedure
takes two columns of untabulated data and
creates several charts for the frequencies of
the unique pairs of values.  Contingency Tables
takes a two-dimensional tabulation and plots it
in various ways.
```



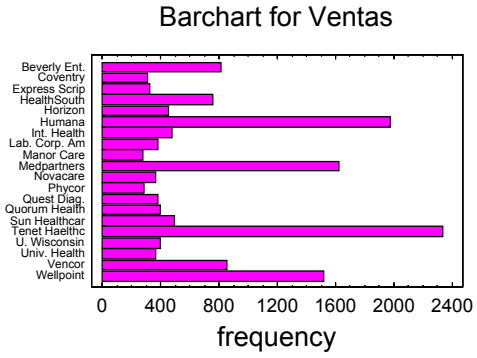
Siguiendo la misma secuencia para el ejemplo 2 de este capítulo, se tienen los siguientes resultados.

Resultados, ejemplo 2

Barchart - Ventas

```
Analysis Summary

Data variable: Ventas
```

```
Sum of 20 values: 14663.0
The StatAdvisor
---------------
   This procedure creates a barchart from a
column of numeric values. A single bar is
plotted for each row of the column. Barcharts
are also created by several other procedures
under Describe - Categorical Data. The
Tabulation procedure will take a column of
untabulated data, find all unique values, and
create a barchart showing how often each unique
values occurs. The Crosstabulation procedure
takes two columns of untabulated data and
creates several charts for the frequencies of
the unique pairs of values. Contingency Tables
takes a two-dimensional tabulation and plots it
in various ways.
```



También se puede trabajar con las posibilidades de **Pane Options** y de **Graphics Options** para darles una mejor presentación a los resultados

La recomendación es trabajar los gráficos desde Statgraphics, para editarlos lo menos posible en el procesador de palabras y evitar que el tamaño de los archivos crezca mucho.

Ejemplo 3

Para "manejar" un poco los datos se pide:
a.- Etiquetar los datos del ejemplo 2 como Año_1977
b.- Generar datos para 1978 con un incremento del 10%
c.- Generar datos para 1979 con un incremento del 15% de los valores de 1978.
d.- Realizar gráficos de barras para describir los datos.

Los datos a trabajar son:

Empresa	Año_1977	Año_1978	Año_1979
Beverly Ent.	805	885.5	1018.325
Coventry	307	337.7	388.355
Express Scrip	320	352	404.8
HealthSouth	748	822.8	946.22
Horizon	445	489.5	562.925
Humana	1968	2164.8	2489.52
Int. Health	472	519.2	597.08
Lab. Corp. Am	377	414.7	476.905
Manor Care	274	301.4	346.61
Medpartners	1614	1775.4	2041.71
Novacare	357	392.7	451.605
Phycor	284	312.4	359.26
Quest Diag.	374	411.4	473.11
Quorum Health	393	432.3	497.145
Sun Healthcar	486	534.6	614.79
Tenet Haelthc	2331	2564.1	2948.715
U. Wisconsin	389	427.9	492.085
Univ. Health	362	398.2	457.93
Vencor	845	929.5	1068.925
Wellpoint	1512	1663.2	1912.68

La secuencia es:

1. Dar un clic sobre el identificador de la columna de ventas y luego un clic derecho, seleccionar **Modify column** y cambiar el nombre.

2. Dar un clic sobre el identificador de la columa 3 (primera fila) y luego un clic derecho. Seleccionar Generate Data, en la caja de diálogo "teclear" o seleccionar la variable para definir la fórmula: Año_1977*1.10 (figura 25)

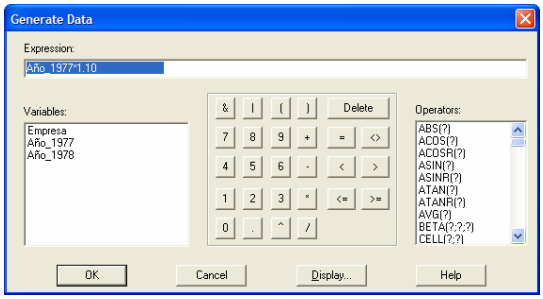


Figura 25. Diálogo para generar nuevos datos.

3. Cambiar el nombre de la columna 3 a Año_1978.

4. Repetir pasos 2 y 3 para la columna 4.

5. Ahora si, del menú seguir la secuencia

Plot -> Business Chart -> Multiple Barchart

6. Colocar las variables adecuadas en la caja de diálogo que aparece, figura 26.

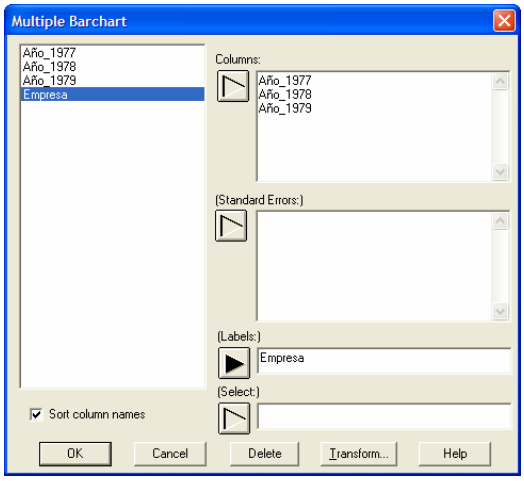


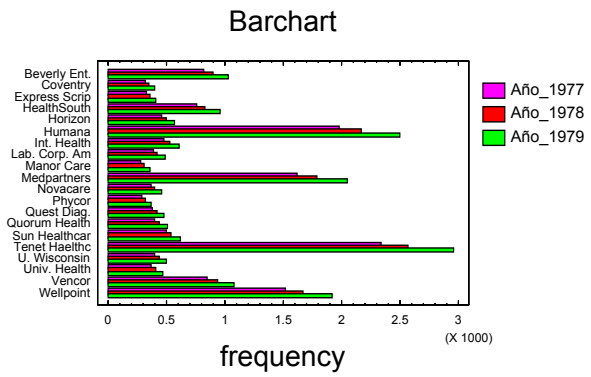
Figura 26. Diálogo para gráficos de barras.

7. Dar un clic sobre el botón OK para obtener los primeros resultados.

Resultados

Multiple Barchart

```
Analysis Summary

Column variables:
     Año_1977
     Año_1978
     Año_1979

Table sum: 49341.0
Number of rows: 20
Number of columns: 3


The StatAdvisor
---------------
   This procedure plots multiple barcharts for
previously tabulated data. If your data has not
yet been tabulated, use the Crosstabulation
procedure instead.
```



Interpretación

Como en los gráficos de barras de una sola variable, este tipo de gráficos sólo permite visualizar quien tiene mayor frecuencia o mayor porcentaje, realizando una comparación relativa entre todas las categorías que se tienen en los datos.

CAPÍTULO 3

DESCRIBIENDO LOS DATOS

En el análisis estadístico la primera actividad a realizar es ver los datos, observarlos o explorarlos. Apreciar su distribución, agrupamiento, dispersión o la presencia de valores extremos. Como se dice: dejar que los datos hablen.

Empezaremos por revisar algunos gráficos de uso común. Como los de Caja y bigote, los de Tallo y hojas, así como los clásicos Histogramas.

DIAGRAMA DE CAJA Y BIGOTE O BOXPLOT

Estas gráficas se han vuelto muy populares, ya que ofrecen mucha información de manera compacta. Muestran el rango de los datos, la dispersión a través del rango intercuartílico y la mediana como medida de tendencia central.

Pasos para construir un BoxPlot

1. Calcular los cuartiles Q_1, Q_2 y Q_3
2. Sobre una línea, horizontal o vertical, "pintar": el valor mínimo; Q_1; Q_2; Q_3 y el valor máximo
3. Hacer un rectángulo de Q_1 a Q_3
4. Trazar una línea en Q_2=mediana
5. Revisar que los valores extremos no estén a una distancia mayor a 1.5 el valor del rango intercuartílico, si hay algún valor marcarlo.

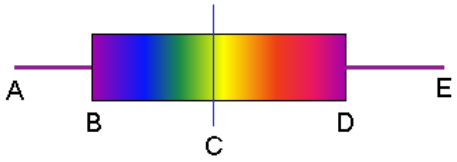


Figura 27. Representación de un Boxplot

Para mostrar su manejo en StatGraphics se consideran los siguientes datos.

Ejemplo 1. En un cierto mes, quince vendedores alcanzaron: 107, 90, 80, 92, 86, 109, 102, 92, 353, 78, 74, 102, 106, 95 y 91 por ciento de sus cuotas de venta.
(Estadística Elemental, John E. Freund y Gary A. Simon, 1992, Prentice Hall, pág. 56).

La secuencia a seguir es:

1. Generar un archivo de datos con una sola columna llamada porcentajes, de tipo numérico.
2. "Teclear" los datos y almacenarlos en disco.
3. Del menú seguir la secuencia

Plot -> Exploratory Plots -> Box-and-Whisker Plot

4. Llenar el diálogo que indica la variable a trabajar, figura 28.

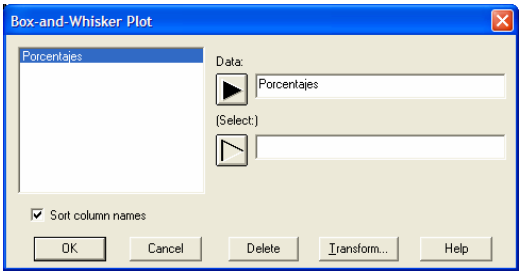


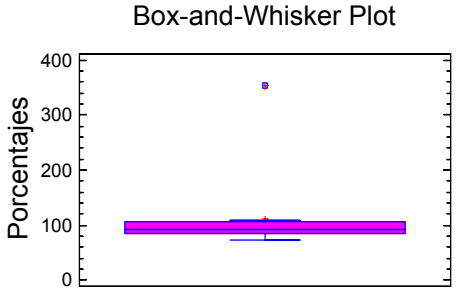
Figura 28. Seleccionar variables para un Boxplot.

5. Presionar el botón OK y analizar los resultados.

Resultados

Box-and-Whisker Plot - Porcentajes

```
Analysis Summary

Data variable: Porcentajes

15 values ranging from 74.0 to 353.0

The StatAdvisor
---------------
   This procedure displays a box-and-whisker
plot of a single column of data. You can create
many other graphs and statistics for the data by
selecting Describe - Numeric Data - One-Variable
Analysis from the main menu.
```



Interpretación

1.- Notar lo concentrado de la caja, en relación sobre todo al alambre de los valores menores.

2.- Notar que la mediana está muy hacia abajo de la caja y no al centro como se espera en una distribución de datos "bien portada".

3.- Nótese el punto aislado, el cual indica un valor extremo.

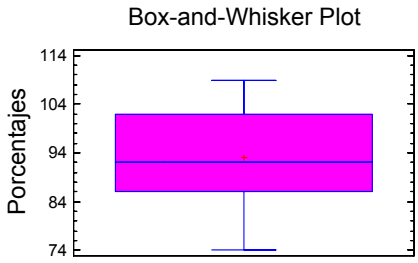
NOTA: El bigote va de la caja al valor más alejado dentro del intervalo de 1.5 rangos intercuartílicos**. Los valores extremos (posibles outliers)**, son casos u observaciones que están entre 1.5 a 3 veces el rango intercuartílico, con respecto al extremo más cercano de la caja. **Outliers o valores aberrantes**, casos con valores más allá de 3 veces el rango intercuartílico, con respecto al extremo más cercano de la caja.

Manejando valores extremos y outliers

Dar un doble clic sobre el gráfico, para maximizarlo. Ahora dar un clic sobre el outlier, para identificar el número de observación y el valor del dato que corresponde a este outlier. Se puede navegar entre ventanas seleccionado del menú la opción **Window** y luego de la parte inferior del menú seleccionar la ventana correspondiente. Generalmente la opción 1 corresponde a los datos, la 4 al StatReporter y los números finales a las ventanas de resultados.

Al Identificar el valor de 353 y eliminarlo se tienen los siguientes resultados.

```
Box-and-Whisker Plot - Porcentajes

Analysis Summary

Data variable: Porcentajes

14 values ranging from 74.0 to 109.0


The StatAdvisor
---------------
   This procedure displays a box-and-whisker
plot of a single column of data. You can create
many other graphs and statistics for the data by
selecting Describe - Numeric Data - One-Variable
Analysis from the main menu.
```



Interpretación

1.- Es importante ver como se modifica la caja y el bigote en este gráfico.

2.- La mediana continúa hacia abajo de la caja y no al centro como se espera en una distribución de datos "bien portada". También es importante notar el signo + dentro de la caja, ya que indica el valor de la media aritmética.

3.- Ya no hay puntos aislados, en otras palabras se eliminó el outlier y se mejoró la distribución de los datos.

NOTA: La identificación de un outlier o valor extremo no implica su eliminación automática. Eliminar datos u observaciones es una decisión que el investigador toma en el contexto de su investigación y no con base en la estadística.

Por otro lado, en el gráfico se pueden analizar las posibilidades del **Pane Options** y de **Graphic Options**, que funcionan de manera semejante a los gráficos de **Pie.**

Ejemplo 2. Un servicio de prueba de consumo obtuvo los siguientes resultados de millas por galón en cinco recorridos de prueba realizados con cada uno de tres automóviles compactos:

Automóvil A	27.9	30.4	30.6	31.4	31.7
Automóvil B	31.2	28.7	31.3	28.7	31.3
Automóvil C	28.6	29.1	28.5	32.1	29.7

(Estadística Elemental, John E. Freund y Gary A. Simon, 1992, Prentice Hall, pág. 56).

Realizar Boxplot´s comparativos de las tres marcas de automóviles.

SOLUCIÓN

1. Primero hay que darle los datos a Statgraphics, nótese el formato en dos columnas: una que identifica el tipo de automóvil y otra para las millas recorridas.

2. Del menú seguir la secuencia:

Plot -> Exploratory Plots ->Multiple Box-and-Whisker Plot

3. Llenar el diálogo que se presenta, figura 29.

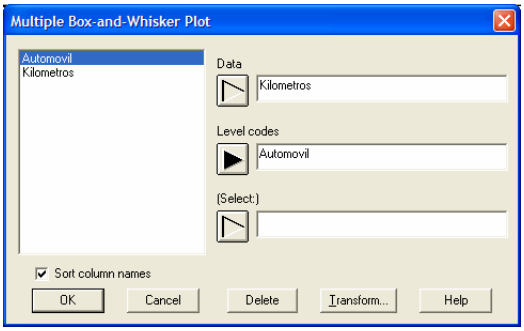
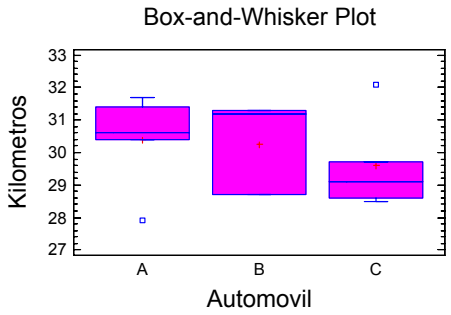


Figura 29. Seleccionar variables para múltiples Boxplot.

4. Dar OK e interpretar los resultados

Multiple Box-and-Whisker Plot - Kilómetros by Automóvil

```
Analysis Summary

Dependent variable: Kilometros
Factor: AutomovilMultiple Box-and-Whisker Plot -
Kilometros by Automovil

Number of observations: 15
Number of levels: 3


The StatAdvisor
---------------
   This procedure constructs box-and-whisker
plots to compare the 3 samples of Kilometros.
For a detailed statistical analysis of this
data, select Compare - Analysis of Variance -
One-Way ANOVA from the main menu.
```



En este gráfico se aprecia la disparidad en la distribución de los datos entre los tres grupos, inclusive se detectan valores extremos en los automóviles A y C.

Otro aspecto relevante es cambiar la dirección de gráfico (horizontal o vertical) con el diálogo **Pane Options** (figura 30) que aparece al dar un clic derecho sobre el gráfico.

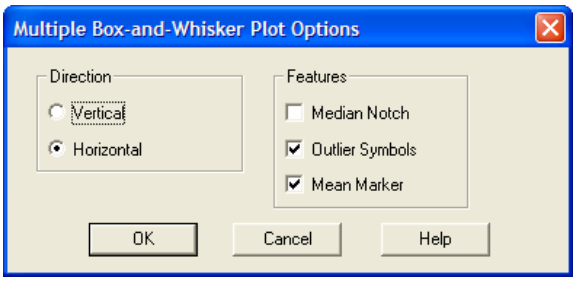
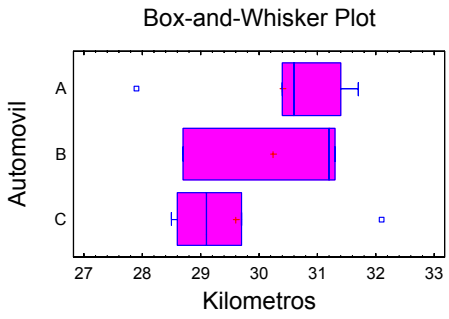


Figura 30. Pane Options para múltiples Boxplot.

Cuyo resultado se presenta a continuación.

Multiple Box-and-Whisker Plot - Kilómetros by Automóvil

```
Analysis Summary

Dependent variable: Kilometros
Factor: Automovil

Number of observations: 15
Number of levels: 3


The StatAdvisor
---------------
This procedure constructs box-and-whisker plots
to compare the 3 samples of Kilometros.  For a
detailed statistical analysis of this data,
select Compare - Analysis of Variance One-Way
ANOVA from the main menu.
```



Un aspecto importante de la interpretación es decir cual de los tres tipos automóvil tiene mejor rendimiento promedio y cuál o cuáles presentan poca dispersión en sus resultados.

NOTA: En estos gráficos, la mediana se representa como una línea vertical dentro de la caja, mientras que la media se indica por un signo de más (+).

DIAGRAMAS DE TALLO Y HOJAS

Estos diagramas fueron desarrollados por John Tukey en 1977. Permiten observar la distribución de los datos originales y son muy útiles para resumir y describir, sobre todo cuando no se rebasan los cien datos.

Para construir un diagrama de tallo y hoja:

1. Colocar a la izquierda los dígitos más significativos del dato (Tallo)

2. Colocar a la derecha los dígitos menos significativos, en orden de menor a mayor, unidades o decimales, (Hojas). En algunos casos conviene poner en las hojas dos dígitos significativos.

3. Hacer un conteo de la frecuencia de valores asociados al valor del tallo.

Y ¿cómo se hacen en Statgraphics?

1. Del menú seguir la secuencia:

Describe -> Numeric Data -> One-Variable Analysis

2. Llenar la información solicitada en la caja de diálogo y seleccionar, en las opciones tabulares de los resultados, el gráfico de tallo y hojas.

El cómo trabajar con estos gráficos se muestra en el ejemplo 3.

HISTOGRAMAS

Un histograma es un gráfico de barras que muestra la frecuencia de cada uno de los valores encontrados en la variable medida (número de veces que se repite un valor). En términos simples, consta de un eje horizontal cuya escala va desde el valor más pequeño hasta el valor máximo en los datos, valores que de preferencia deben ser cuantitativos y continuos (resultado de mediciones de peso, longitud, volumen, etc.); y de un eje vertical cuya escala puede ir desde cero hasta la máxima frecuencia encontrada.

Pasos para construir un histograma.

Para elaborar un histograma se recomienda:

1) Obtener el valor máximo y el mínimo, de todo el conjunto de valores.

2) Escribir cada uno de los valores, en columna y en orden ascendente.

3) Revisar todos los valores del conjunto total de datos y colocar una marca, al frente de cada valor, por cada vez que se repita.

4) Contar el número de marcas en cada uno de los valores, anotándolo en la fila correspondiente.

Y ¿cómo se hacen en Statgraphics?

Seguir la misma secuencia de los gráficos de tallo y hojas. En las opciones tabulares aparece la tabulación de frecuencias y en opciones gráficas se puede solicitar el histograma de frecuencias.

Ejemplo 3, (considerando los datos del ejemplo 1 de este capítulo, eliminando el outlier)

Realizar un gráfico de tallos y hojas; un histograma y empezar a describir los datos.

1. "Teclear" los datos en una columna llamada porcentaje, de tipo numérico.

2. Seguir la secuencia

Describe -> Numeric Data -> One-Variable Analysis

3. En la caja de diálogo seleccionar la variable a analizar

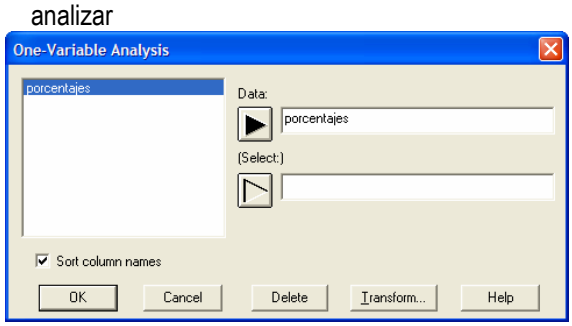


Figura 31. Diálogo para el Análisis de una variable.

4. Dar OK para visualizar los primeros resultados
5. Dar un clic sobre las opciones tabulares (**Tabular Options**, ícono en el botón color amarillo de la ventana de resultados), figura 32.

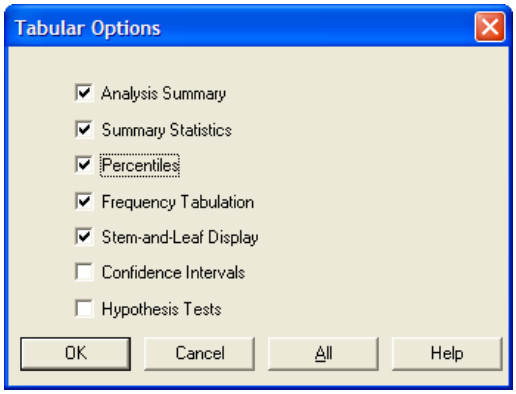


Figura 32. Opciones tabulares para el Análisis de una variable.

Seleccionar, todo menos intervalos de confianza y prueba de hipótesis que corresponde a la inferencia estadística.

6. Dar un clic sobre las opciones gráficas (**Graphical Options**, ícono en el botón color azul de la ventana de resultados), figura 33.

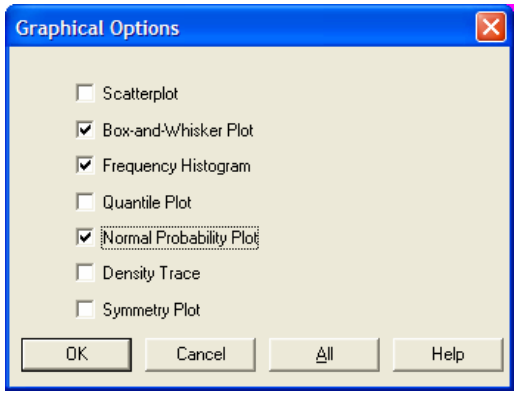
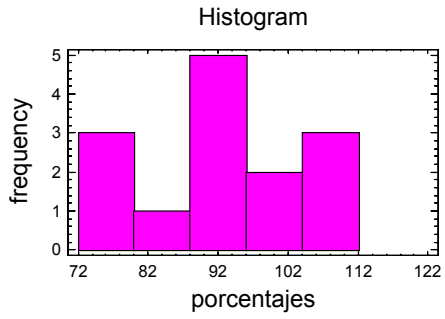


Figura 33. Opciones gráficas para el Análisis de una variable.

7. Seleccionar gráficos de Boxplot, Histogramas de frecuencia y los gráficos de Probabilidad Normal.

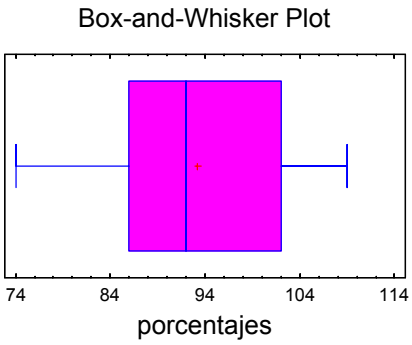
8. Dar OK para observar los resultados.



Resultados

One-Variable Analysis - porcentajes

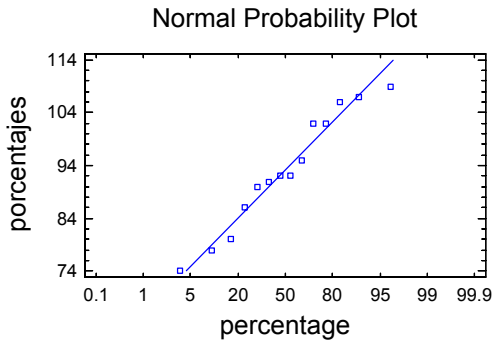
```
Analysis Summary

Data variable: porcentajes

14 values ranging from 74.0 to 109.0

The StatAdvisor
---------------
   This procedure is designed to summarize a
single sample of data. It will calculate various
statistics and graphs. Also included in the
procedure are confidence intervals and
hypothesis tests. Use the Tabular Options and
Graphical Options buttons on the analysis
toolbar to access these different procedures.
```




```
Summary Statistics for porcentajes

Count = 14
Average = 93.1429
Variance = 123.516
Standard deviation = 11.1138
Minimum = 74.0
Maximum = 109.0
Range = 35.0
Stnd. skewness = -0.29921
Stnd. kurtosis = -0.733076


The StatAdvisor
---------------
   This table shows summary statistics for
porcentajes.  It incluyes measures of central
tendency, measures of variability, and measures
of shape.  Of particular interest here are the
standardized skewness and standardized kurtosis,
which can be used to determine whether the
sample comes from a normal distribution.  Values
of these statistics outside the range of -2 to
+2 indicate significant departures from
normality, which would tend to invalidate any
statistical test regarding the standard
deviation.  In this case, the standardized
skewness value is within the range expected for
data from a normal distribution.  The
standardized kurtosis value is within the range
expected for data from a normal distribution.
```

```
Percentiles for porcentajes

1.0% = 74.0
5.0% = 74.0
10.0% = 78.0
25.0% = 86.0
50.0% = 92.0
75.0% = 102.0
90.0% = 107.0
95.0% = 109.0
99.0% = 109.0
```

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



Este gráfico permite apreciar las desviaciones de la normalidad de los datos. Para ser normales deben comportarse linealmente.

```
Frequency Tabulation for porcentajes

-----------------------------------------------------------------
      Lower Upper          Relative  Cumulative Cum.       Rel.
Class Limit Limit Midpoint Frequency Frequency  Frequency  Freq
-----------------------------------------------------------------
at or below  72.0             0       0.0000       0      0.0000
1     72.0   80.0   76.0      3       0.2143       3      0.2143
2     80.0   88.0   84.0      1       0.0714       4      0.2857
3     88.0   96.0   92.0      5       0.3571       9      0.6429
4     96.0  104.0  100.0      2       0.1429      11      0.7857
5    104.0  112.0  108.0      3       0.2143      14      1.0000
above 112.0                   0       0.0000      14      1.0000
-----------------------------------------------------------------
Mean = 93.1429   Standard deviation = 11.1138
```

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

```
Stem-and-Leaf Display for porcentajes: unit =
1.0   1|2 represents 12.0

      1     7|4
      2     7|8
      3     8|0
      4     8|6
     (4)    9|0122
      6     9|5
      5    10|22
      3    10|679
```

```
The StatAdvisor
---------------
   This display shows a frequency tabulation for
porcentajes. The range of the data has been
divided into 8 intervals (called stems), each
represented by a row of the table. The ítems
are labeled using one or more leading digits for
the data values falling within that interval.
On each row, the individual data values are
represented by a digit (called a leaf) to the
right of the vertical line. This results in a
histogram of the data from which you can recover
at least two significant digits for each data
value. If there are any points lying far Hawai
from most of the others (called outside points),
they are placed on separate high and low stems.
In this case, there are no outside points.
Outside points are illustrated graphically on
the box-and-whisker plot, which you can access
via the list of Graphical Options. The leftmost
column of numbers are depths, which give
cumulative counts from the top and bottom of the
table, stopping at the row which contains the
median.
```

Interpretación.

Este software presenta un **StatAdvisor,** herramienta de interpretación que sólo requiere leer en ingles.

También aparecen las medidas de tendencia central y de dispersión como la media, mediana, moda, varianza, desviación estándar, rango, sesgo y curtosis estándares.

El gráfico Boxplot que da una idea de la distribución de los datos, además del gráfico de tallo y hojas, que muestra la distribución de frecuencia de los datos y donde se pueden apreciar los datos originales.

Además, se tiene una tabla de frecuencias, cuyo gráfico también se muestra.

El gráfico de probabilidad normal permite visualizar hasta donde los datos se apegan a una distribución normal o se alejan de ella, aspecto interesante y necesario para trabajar la estadística inferencial.

Más opciones de análisis

Statgraphics muestra resultados parciales en ventanas separadas, cada una de las cuales puede tener a su vez más opciones.

Para accesar a estas opciones hay que dar un clic derecho sobre alguna ventana de resultados y del menú flotante que aparece seleccionar **Pane Options**. Por ejemplo para **Summary** se presenta el diálogo de la figura 34.

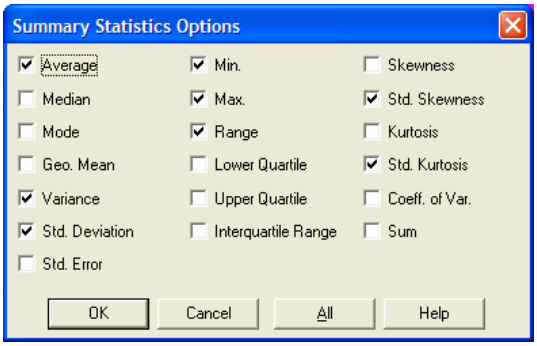


Figura 34. Estadísticas básicas para el Análisis de una variable.

Aquí basta con dar un clic sobre la caja de selección de aquellas opciones que se requieran (de preferencias las que sepa Ud. interpretar).

En la ventana de frecuencias se tiene un diálogo que permite definir el número de intervalos de clases, así como el límite inferior y superior de los datos (figura 35). Este diálogo también afecta al gráfico de histograma de frecuencias.

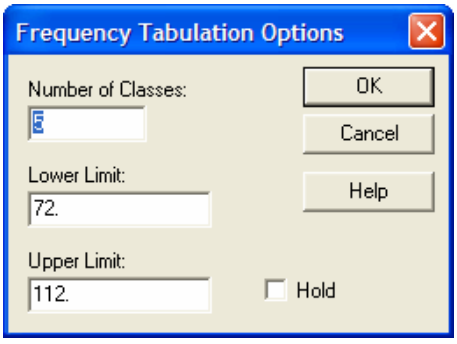


Figura 35. Opciones para tabla de frecuencias.

Las gráficas también tienen más opciones, en el Pane Options. Por ejemplo, para el histograma de frecuencias se presenta el diálogo de la figura 36.

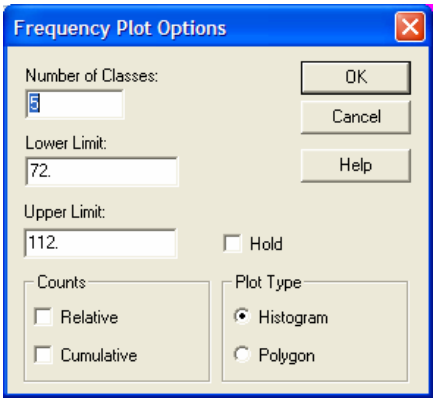


Figura 36. Opciones para el gráfico de frecuencias.

NOTA: Si se selecciona Polygon + Cumulative se obtiene la Ojiva.

Ejemplo 4

Se tienen los siguientes datos de emisiones de óxido de azufre, en toneladas.
(modificado de Estadística Elemental, John E. Freund y Gary A. Simon, 1992, Prentice Hall, pp. 21-22).

Planta industrial A

15.8	22.7	26.8	19.1	18.5	14.4	8.3	25.9	26.4	9.8
22.7	15.2	23.0	29.6	21.9	10.5	17.3	6.2	18.0	22.9
24.6	19.4	12.3	15.9	11.2	14.7	20.5	26.6	20.1	17.0
22.3	27.5	23.9	17.5	11.0	20.4	16.2	20.8	13.3	18.1

Planta industrial B

27.8	29.1	23.9	24.4	21.0	27.3	14.8	20.9	21.7	15.8
18.5	22.2	10.7	25.5	22.3	12.4	16.9	31.6	22.4	24.6
16.5	27.6	23.0	27.1	12.0	20.6	19.7	19.9	26.5	21.4
28.7	23.1	16.2	26.7	13.7	22.0	17.5	21.1	34.8	31.5

a. "Teclear" y guardar los datos en un archivo Statgraphics (.sf3)

b. Realizar un análisis descriptivo de los datos y comparar las dos plantas.

c. Guardar los resultados del análisis en un archivo Word

SOLUCIÓN:

a. Generar un archivo con dos columnas, ambas de tipo numérico, una para el tipo de planta (1 o 2) y otra para las toneladas de azufre; con 80 filas o renglones de datos.

Asegurarse de almacenar el archivo en disco duro o flexible.

b. Para realizar el análisis, del menú seguir la secuencia
Describe -> Numeric Data -> Subset Analysis

En el diálogo que aparece ubicar las variables en su lugar correspondiente.

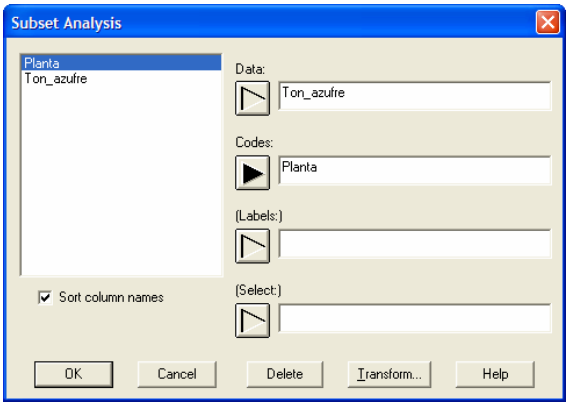
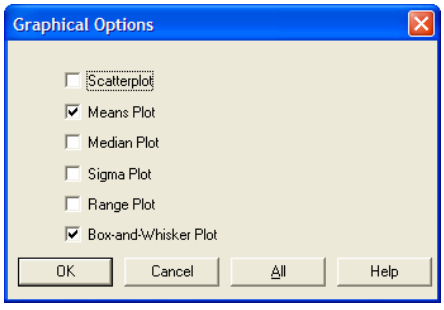


Figura 37. Opciones para Subset Análisis.

Data corresponde a los datos que se quieren analizar y **Code** a la variable que identifica las categorías o subconjuntos de datos.

Dar un clic sobre el botón OK para visualizar los resultados parciales.

Después seleccionar las opciones tabulares y las opciones gráficas de este análisis.



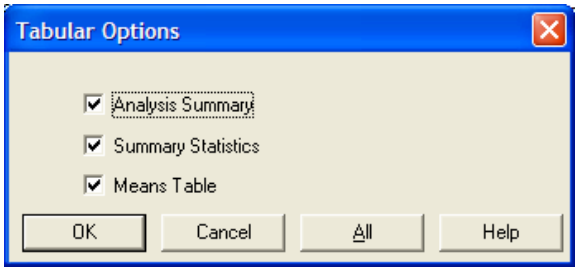
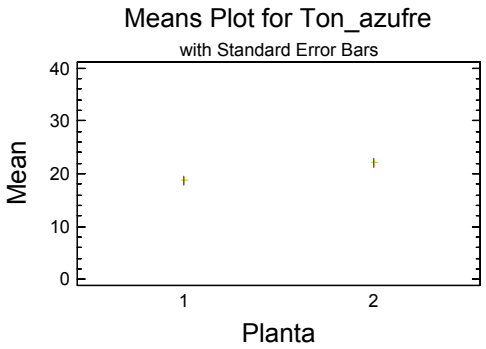


Figura 38. Opciones tabulares y gráficas de Subset Análisis.

Resultados

Subset Analysis

```
Analysis Summary

Data variable: Ton_azufre
Code variable: Planta

Number of observations: 80
Number of levels: 2


The StatAdvisor
---------------
This procedure calculates summary statistics for
the values of Ton_azufre corresponding to each
of the 2 levels of Planta.  It also creates a
variety of plots and allows you to save the
calculated statistics.  Further analyses can be
performed on the data using the Oneway Analysis
of Variance procedure under Compare on the main
menu.
```

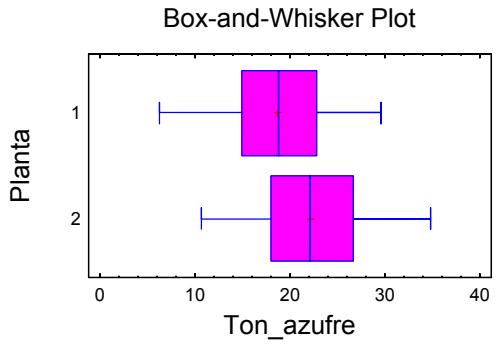



```
Summary Statistics
```

Code	Count	Average	Median	Standard Deviation	Minimum	Maximum	Range
1	40	18.7075	18.8	5.71697	6.2	29.6	23.4
2	40	22.085	22.1	5.66168	10.7	34.8	24.1
Total	80	20.3963	20.85	5.90317	6.2	34.8	28.6

Code	Interquartile Range	Skewness	Kurtosis
1	7.85	-0.196218	-0.584771
2	8.6	-0.00843432	-0.34117
Total	8.3	-0.0971557	-0.334444

```
The StatAdvisor
---------------
This table shows sample statistics for the 2
levels of Planta.
```



Means Table with Standard Error Intervals

Code	Count	Mean	Standard Error	Lower Limit	Upper Limit
1	40	18.7075	0.903933	17.8036	19.6114
2	40	22.085	0.89519	21.1898	22.9802
Total	80	20.3963	0.659995	19.7363	21.0562

The StatAdvisor

This table shows the sample means and Standard errors for the 2 levels of Planta. Also show are intervals representing the means plus and minus one standard error.

c. Para almacenar los resultados se debe dar un clic derecho en cualquier ventana de resultados y del menú flotante que aparece elegir **Copy Analysis to StatReporter.** Después se puede salvar el StatReporter como un archivo Word (**File -> Save as -> Save StatReporter as**) o simple y sencillamente ir a la ventana de StatReporter (**Window -> StatReporter**) y hacer un Copia y Pega o un Corte y Pega de toda la información de interés.

Interpretación

Se puede apreciar que la emisión de azufre es, en promedio, semejante en ambas plantas (tanto en la media como en dispersión). Aunque la planta B tiende a presentar una mayor emisión de Azufre.

NOTA: **Todas las herramientas descriptivas muestran su verdadera utilidad cuando se tienen dos o más grupos a comparar, por lo que es interesante analizar comparativamente: medias, varianzas, rangos, sesgos y curtosis; tanto a nivel numérico como de manera gráfica**.

Ejemplo 5. Retomando el ejemplo del servicio de prueba de consumo donde se obtuvieron los siguientes resultados de milla por galón en cinco recorridos de prueba realizados con cada uno de tres automóviles compactos:

Automóvil A	27.9	30.4	30.6	31.4	31.7
Automóvil B	31.2	28.7	31.3	28.7	31.3
Automóvil C	28.6	29.1	28.5	32.1	29.7

Realizar el análisis descriptivo correspondiente.

Aquí se tiene un archivo de datos con 2 columnas y 15 datos. Una de nombre marca y la otra recorrido.

1. A partir de la barra de menú seguir la secuencia

Describe -> Numeric Data -> Subset Analysis

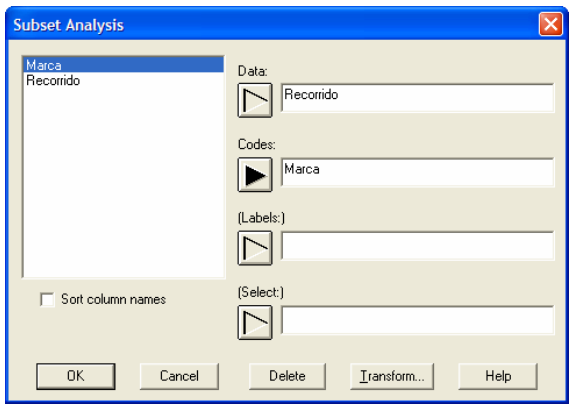


Figura 39. Diálogo para seleccionar variables.

2. Una vez que se seleccionan las variables dar OK y observar los resultados.

3. Considerar las opciones tabulares y gráficas, totalmente análogas a cuando se trabajó con un solo grupo

4. Revisar las posibilidades de **Pane Options**, en cada ventana de resultados.

5. Resultados

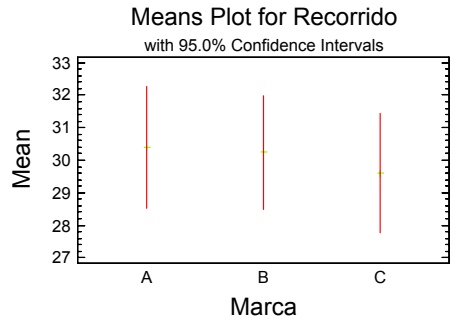
Subset Analysis

Analysis Summary
Data variable: Recorrido
Code variable: Marca
Number of observations: 15
Number of levels: 3

The StatAdvisor

This procedure calculates summary statistics for the values of Recorrido corresponding to each of the 3 levels of Marca. It also creates a variety

of plots and allows you to save the calculated statistics. Further analyses can be performed on the data using the Oneway Analysis of Variance procedure under Compare on the main menu.

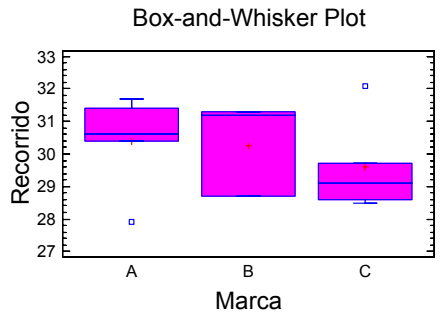


Summary Statistics

Code	Count	Average	Median	Standard Deviation
A	5	30.4	30.6	1.49833
B	5	30.24	31.2	1.40641
C	5	29.6	29.1	1.47648
Total	15	30.08	30.4	1.39908

The StatAdvisor

This table shows sample statistics for the 3 levels of Marca.



Means Table with 95.0% Confidence Intervals

Code	Count	Mean	Standard Error	Lower Limit	Upper Limit
A	5	30.4	0.670075	28.5396	32.2604
B	5	30.24	0.628967	28.4937	31.9863
C	5	29.6	0.660303	27.7667	31.4333
Total	15	30.08	0.361241	29.3052	30.8548

The StatAdvisor

This table shows the sample means and standard errors for the 3 levels of Marca. Also shown are intervals representing 95.0% confidence intervals for the means.

NOTA: El nivel de confianza se puede modificar, para esto dar un clic derecho en la ventana de resultados, seleccionar **Pane Options** y en el diálogo que aparece modificar el valor del nivel de confianza.

¿Qué puede interpretar de estos resultados?

Ejercicios

Seguir toda la secuencia de análisis, más algunas que Ud. explore y considere que mejoran y facilitan el análisis estadístico de los siguientes datos.

1. En el primer día de clases se les preguntó a 50 estudiantes acerca del tiempo requerido para desplazarse de su casa a la universidad (redondeado a 5 minutos). Los datos resultantes son:

30	40	35	50	50	55	50	35	35	60
45	45	20	50	55	70	45	60	30	45
35	50	20	45	75	55	55	45	35	30
25	40	25	45	50	55	40	40	40	40
15	40	25	35	35	50	50	30	50	35

¿Qué se puede decir de los tiempos de desplazamiento? Sugerencia analizar con base en dispersiones y medidas de tendencia central.

2. Se toma una muestra de 50 calificaciones de una población de resultados de un examen final de estadística. Estos datos se muestran en la siguiente tabla.

75	97	71	65	84	27	99	91	99	82
96	58	94	43	10	10	91	10	94	43
74	73	68	54	50	49	81	10	97	76
10	94	79	80	82	71	88	88	47	73
71	99	86	10	84	93	77	98	44	10

¿Qué tan buenas son las calificaciones y qué se puede decir de este grupo de alumnos?

SUGERENCIA. Como sólo hay una variable, se recomienda seguir la secuencia

Describe -> Numeric Data -> One-Variable Analysis

REVISIÓN DE CONCEPTOS BÁSICOS

Para entender e interpretar los resultados obtenidos hasta el momento, se revisan algunos conceptos estadísticos muy sencillos pero útiles.

Tipo de Datos y Niveles de Medición

Los datos pueden ser cualitativos o cuantitativos. Los **datos cualitativos**, como color de ojos en grupo de individuos, no se pueden trabajar aritméticamente, ya que son etiquetas que definen si un individuo pertenece o no a una categoría. Inclusive los datos de este tipo también se conocen como categóricos.

Datos cuantitativos: Mediciones que toman valores numéricos, para los cuales descripciones como media y desviación estándar son significativos. Se pueden clasificar en discretos y continuos.

Datos discretos: Se colectan por **conteo**, por ejemplo el número de productos defectuosos en un lote de producción.

Datos continuos: Se colectan por **medición** y se expresan en una escala continua. Por ejemplo la estatura de las personas.

La primera actividad en el análisis estadístico es contar o medir, ya que los datos representan un modelo de la realidad basado en escalas numéricas y medibles. Los datos vienen en forma Nominal, Ordinal, Intervalo y de razón o cociente.

Los datos categóricos se pueden medir sobre una escala nominal u ordinal. Mientras que en los datos continuos se pueden medir en escala de intervalo o de razón. En la escala de intervalo el valor cero y las unidades de medición son arbitrarias, mientras que en la escala de razón la unidad de medición es arbitraria pero el cero es una característica natural.

Considerando que la mayoría de las pruebas estadísticas clásicas se basan en que los datos se apeguen a una distribución normal, se prefieren los datos en escalas de intervalo o de razón.

Medidas de Tendencia Central

Una forma de representar a un conjunto de datos es a través de una medida de localización de la tendencia central, entre las cuales se tiene la media, mediana y moda.

Media. La media aritmética o simplemente promedio se obtiene sumando todos los valores de una muestra y dividiendo entre el número de datos.

$$Media = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

NOTA: Se considera el valor de Y, por corresponder a la variable de respuesta, que es la que se mide y para no confundir con X´s o variables independientes que en muchos estudios son factores o variables categóricas.

El cálculo de la media utiliza todas las observaciones, de manera que se vuelve sensible a valores extremos, ya que valores extremadamente grandes o pequeños "jalan" el resultado de cálculo hacia ellos. Aún así, es la medida de tendencia central más utilizada, por presentar propiedades matemáticas convenientes para el análisis estadístico inferencial.

Media ponderada. Se utiliza cuando los datos de una muestra no tienen todos los mismos pesos, entonces cada valor se pondera por un peso acorde a su nivel de importancia.

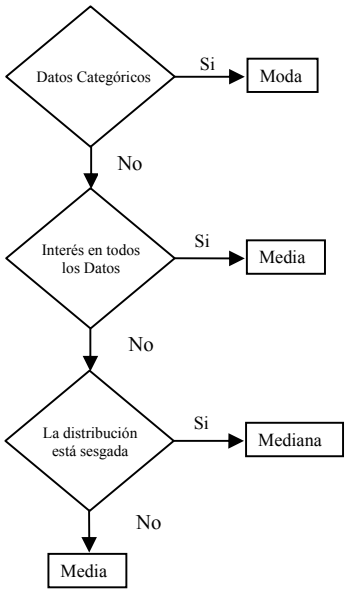
$$\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$$

Mediana. Es el valor que está exactamente a la mitad de un conjunto ordenado de observaciones. Si el número de observaciones es impar la mediana es un sólo valor. Si el número de observaciones es par entonces la mediana es el promedio de los dos valores que están al centro.

Generalmente la mediana es una mejor medida de tendencia central cuando hay valores extremadamente pequeños o grandes y cuando los datos están sesgados a la izquierda o la derecha.

Moda. La moda es el valor que más se presenta en un conjunto de observaciones.

Seleccionando entre media, mediana y moda



Medidas de Dispersión

Al hablar de dispersión se debe considerar que la calidad de la información y la variación están inversamente relacionadas. De aquí la necesidad de medir la variación que existe en un conjunto de datos.

Las medidas más comunes de variación son: el rango, varianza, desviación estándar y coeficiente de variación.

Rango. Es el valor absoluto de la diferencia del valor máximo menos el valor mínimo. Sólo se basa en dos valores y no es una medida recomendable cuando hay valores extremos.

R = Valor máximo – Valor mínimo

$$R = y_n - y_1$$

Cuando se trabaja con datos discretos es común definirlo como:

$$R = y_n - y_1 + 1.$$

Cuartiles. Cuando se tienen un conjunto de datos ordenados en forma ascendente, se pueden dividir en cuartos, Q_1, Q_2, Q_3 y Q_4. Para el valor del primer cuartil, Q_1, hay un 25% de valores más pequeños y un 75% de valores más grandes, de manera análoga en el Q_2=mediana hay un 50% de valores más pequeños y un 50% de valores más grandes y en Q_3 un 75% y 25% respectivamente.

Percentiles. Es un concepto semejante al de cuartil, de tal manera que $Q_1 = P_{25}$, $Q_2 = P_{50}$ = mediana y $Q_3 = P_{75}$. La ventaja de los percentiles es que pueden dividir a un conjunto de datos en 100 partes.

Rango intercuartilico. Expresa el intervalo de valores en el cual se encuentra el 50% de los datos, ya que es la distancia del cuartil 1 al cuartil 3, esto es:

$$RIQ = Q_3 - Q_1$$

Varianza. Es un promedio de las distancias de cada observación con respecto a la media aritmética.

$$Varianza = s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

La varianza es una medida de la variabilidad o dispersión de los datos y no está en las mismas unidades que las observaciones, de ahí que sea difícil su interpretación. Este problema se resuelve trabajando con la raíz cuadrada de la varianza.

Desviación estándar = s = $\sqrt{s^2}$

Coeficiente de variación. Expresa la variación de un conjunto de datos en relación a su media.

$$CV = 100\left|\frac{s}{\bar{y}}\right|\%$$

El CV es independiente de las unidades de medición y en la estimación de un parámetro, cuando es menor al 10% el estimador se considera aceptable. Al inverso del CV, 1/CV, se le conoce como el cociente señal/ruido.

Para datos sesgados o agrupados, el coeficiente de variación cuartil es más útil que el CV.

$$V_Q = \left(\frac{Q_3 - Q_1}{Q_3 + Q_1}100\right)\%$$

Sesgo y Curtosis

Sesgo. Es una medida de la desviación de una muestra con respecto a la media de una distribución normal. En otras palabras, mide la asimetría en la distribución de un conjunto de datos.

El sesgo es cero cuando se tiene una distribución simétrica con respecto a la media. Cuando es positivo indica que las observaciones se agrupan a la izquierda de la media, con la mayoría de los valores extremos a la derecha de la media. En otras palabras el signo del sesgo indica hacia que lado de la media se tienen los valores extremos.

$$\text{Sesgo (Skewness)} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^3}{(n-1)s^3}, \text{ con } n > 1$$

Curtosis. Es una medida del pico o aplanado de una distribución. Una distribución normal estándar tiene una curtosis de 3. De tal manera que un valor mayor que 3.0 indica un pico mayor a una distribución normal, mientras un valor menor que 3.0 indica una distribución más aplanada que una normal.

$$\text{Curtosis} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^4}{(n-1)s^4}, \text{ con } n > 1$$

Outlier

Un outlier u observación aberrante es un resultado distante de la mayoría de las observaciones. Se identifica porque su distancia al cuartil más cercano es mayor a 1.5 veces el rango intercuartílico.

La estadística, como toda disciplina científica, tiene su propio lenguaje o "jerga técnica" que permite y facilita la comunicación entre estadísticos o entre quien tiene los datos y quien debe analizarlos, de ahí la necesidad de definir y conocer algunos términos o conceptos.

GLOSARIO

Población: Colección de personas, animales, plantas o cosas acerca de las cuales se colectan datos. Es el grupo de interés sobre el cual se quieren realizar conclusiones.

Variables cualitativas y cuantitativas: Cualquier objeto o evento, que puede variar en observaciones sucesivas, ya sea en cantidad o cualidad. De aquí que se clasifiquen como cuantitativas o cualitativas, cuyos valores se denominan "variedades" y "atributos" respectivamente.

Variable: Una característica o fenómeno, que toma valores diferentes, como: peso o género, ya que difieren de medición a medición.

Aleatoriedad: esto significa impredecibilidad. Lo fascinante es que aunque cada observación aleatoria puede no ser predecible por si sola, colectivamente siguen un patrón predecible, llamado función de distribución. Lo que permite asociarle una probabilidad a la ocurrencia de cada resultado.

Muestra: Un subconjunto perfectamente acotado y definido de una población o universo.

Experimento: Es un proceso cuyos resultados no se conocen de antemano ni son predecibles.

Experimento estadístico: En general un experimento es una operación en la cual se seleccionan o fijan los valores de una variable (variable independiente) y se miden o cuantifican los valores de otra variable (variable dependiente). Entonces, un experimento estadístico es una operación en la cual se fijan los valores de la variable independiente y se toma una muestra aleatoria de una población para inferir los valores de la variable dependiente.

Diseño de Experimentos: Es una herramienta para adquirir conocimiento acerca de un fenómeno o proceso. Este conocimiento se puede utilizar para ganar competitividad, acortar el ciclo de desarrollo de un producto o proponer nuevos productos o procesos que cumplan o excedan la expectación de un comprador.

Variable aleatoria: Es una función real (se le llama variable pero en realidad es una función) que asigna un valor numérico a cada evento simple. Estas variables son necesarias ya que no se pueden realizar operaciones algebraicas sobre resultados textuales, lo que permite obtener estadísticas, como promedios y varianzas. Además de que cualquier variable aleatoria tiene asociada una distribución de probabilidades.

Probabilidad: En términos simples es una medida de la posibilidad, se puede definir como el cociente del número de casos de interés que se presentan en un estudio entre el número de casos totales. Se utiliza para

anticipar el tipo de distribución que sigue un conjunto de datos y asociarlos a un modelo probabilístico. Es importante hacer notar que los fenómenos aleatorios no son azarosos, ya que presentan un orden que sólo emerge cuando se describe un gran número de corridas (por ejemplo, al lanzar dos veces una moneda rara vez se obtiene un sol y una águila, pero si la lanza digamos unas diez mil veces, lo más seguro es que exista una clara tendencia a obtener la mitad de lanzamientos como sol y la otra mitad como águila). La descripción matemática de la variación es central a la estadística, ya que la probabilidad requerida para la inferencia está orientada hacia la distribución de los datos y no es de ninguna manera axiomática o combinatoria.

Unidad Muestral: Es una persona, animal, planta o cosa que está bajo observación o estudio por un investigador. En otras palabras, el objeto o entidad básica sobre la cual se realiza un experimento, por ejemplo, una persona, una muestra de suelo, etcétera.

Parámetro: Un valor desconocido, que por lo tanto tiene que estimarse. Los parámetros se utilizan para representar alguna característica de una población. Por ejemplo, la media poblacional, μ, es un parámetro que generalmente se utiliza para indicar el promedio de una característica numérica de la población.

En una población, un parámetro es un valor fijo que no cambia. Cada muestra de la población tiene su propio valor de alguna estadística que sirve para estimar su parámetro.

Estadística o estadístico: Valor que se obtiene a partir de una muestra de datos. Se utiliza para generar información acerca del parámetro de su población. Ejemplo de estadística es la media y la varianza de una muestra, que se representan con letras latinas, mientras que los parámetros de una población se representan con letras griegas.

Estadística descriptiva: Área de la estadística que permite presentar los datos de manera clara y concisa, de tal forma que para la toma de decisiones se tengan a la mano las características esenciales de los datos.

Los principales estadísticos como medidas de tendencia central son la media o la mediana, y la varianza o la desviación estándar como medidas de dispersión.

Estadística inferencial. Implica o significa hacer inferencias de una población, partiendo de valores muestrales. Cualquier conclusión de este tipo se debe expresar en términos probabilísticos, ya que la probabilidad es el lenguaje y la herramienta de medición de la incertidumbre que cuantifica la validez de una conclusión estadística.

Inferencia estadística: Se refiere a incrementar el conocimiento de una población con base en datos muestrales, también se conoce como razonamiento inductivo e implica conocer el todo a partir de sus partes. La inferencia estadística guía la selección de un modelo estadístico apropiado que en combinación con el adecuado tipo de datos permite cuantificar la confiabilidad de las conclusiones que se obtienen.

Condiciones de Distribución normal: La distribución normal o Gaussiana es una distribución continua y simétrica que tiene la forma de una campana. Uno de los aspectos más interesantes es que sólo se necesitan la media y la varianza para determinar completamente la distribución. En estudios reales se han encontrado una amplia gama de variables que tienen una distribución aproximadamente normal. Y que aún cuando una variable puede seguir una distribución no-normal la media de muchas observaciones independientes de la misma distribución se acerca arbitrariamente a una distribución normal, conforme el número de observaciones aumenta.

Su mayor importancia radica en que muchas de las pruebas estadísticas más frecuentemente utilizadas tienen como condición que los datos sigan una distribución normal.

Estimación y contraste o prueba de hipótesis: La inferencia en estadística considera dos grandes temas: el primero es la estimación, que implica estimar valores para parámetros poblacionales, considerando la variación aleatoria, por lo que dichas estimaciones se dan en valores de intervalo y nunca como valores puntuales. El segundo gran tema corresponde al contraste o prueba de hipótesis, donde se someten a prueba los posibles valores de algún parámetro con base en un modelo estadístico probabilístico.

COMENTARIOS FINALES

Hasta el momento se han revisado los aspectos básicos del "paquete" Statgraphics y del análisis estadístico, por lo que estamos listos para empezar a revisar los fundamentos de la estadística inferencial, es decir estimaciones y contrastes o pruebas de hipótesis.

CAPÍTULO 4

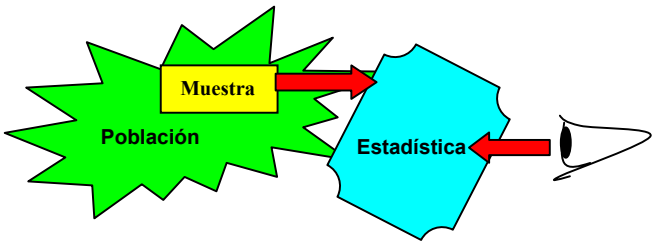
INFERENCIA ESTADÍSTICA

MOTIVACIÓN

La inferencia estadística se caracteriza porque a través de una muestra se pueden realizar inferencias de toda una población en estudio. De manera que utilizando modelos estadísticos se puede asignar un nivel de confiabilidad a las conclusiones que se obtengan, proporcionando soporte para la toma de decisiones.

Población y muestra

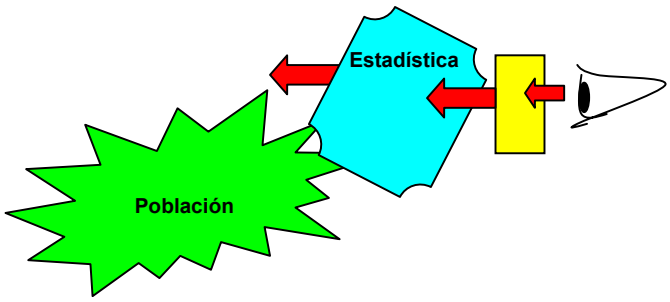
En cualquier proceso de investigación o producción es demasiado costoso, en recursos o en tiempo, revisar uno a uno todos los elementos que conforman una población, de ahí la necesidad de revisar unos cuantos, que sean representativos, y a partir de ellos predecir el comportamiento de toda la población.



El primer "viaje" a la estadística implica seleccionar una muestra de manera aleatoria, es decir, sin privilegiar o descartar de antemano elemento alguno; garantizando que todos tengan la misma posibilidad de ser elegidos. La mejor forma de hacer esto es utilizando herramientas como tablas de números aleatorios, una urna, o algún proceso de números pseudoaleatorios como los que vienen integrados las calculadoras y en la mayoría de los paquetes estadísticos. Cualquiera de estas opciones es mejor que cerrar los ojos y estirar la mano o establecer criterios personales de selección de muestras.

Uno de los ejemplos más simples, pero nada estadístico, es lo que hacen quienes cocinan ya que a través de pequeñas "probadas" saben si un guiso está o no en su punto, esto previa homogenización del contenido de la cazuela y sin consumir todo su contenido.

Es conveniente aclarar que el tema de muestreo es una de las grandes ramas de la estadística, para la cual existen libros completos que analizan a detalle cada una de las opciones, dependiendo del propósito del muestreo.



El segundo "viaje" a la estadística consiste en analizar la muestra mediante alguna de las muchas técnicas de la estadística inferencial para tomar decisiones con respecto a la población, apoyándose en el conocimiento de causa evidenciado a partir de los datos y asignándole un nivel de confiabilidad o de incertidumbre a las conclusiones obtenidas.

INCERTIDUMBRE Y DISTRIBUCIONES ESTADÍSTICAS

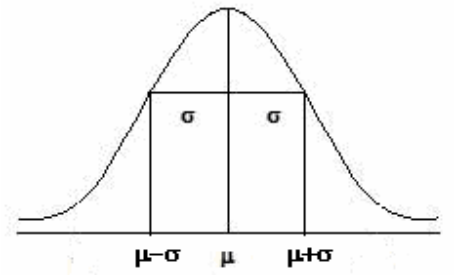
La estadística es la disciplina que estudia los procesos estocásticos, es decir aquellos que presentan variaciones, sin causa asignable (debidas al azar). Por lo que se han desarrollado técnicas que permiten detectar y diferenciar variaciones por efecto de algún factor, de las debidas al azar, con el fin de identificar su comportamiento y reducir estas últimas a un nivel aceptable para que no altere las características de calidad de los productos en manufacturación.

Con el apoyo de la teoría de la probabilidad se ha demostrado que las variables aleatorias tienen un comportamiento bien definido, que se puede representar mediante funciones de probabilidad y funciones de densidad de probabilidad, que dependiendo del tipo de unidades de medición generan las distribuciones estadísticas, base fundamental de las técnicas inferenciales. Debido a su importancia algunas de ellas se han tabulado para facilitar su uso; entre las más conocidas, sin ser las únicas, se encuentran:

- Binomial
- Poisson
- Normal (Z)
    - t-student
    - F-Fisher
    - Ji-cuadrada (χ^2)

Estas distribuciones realmente corresponden a modelos matemáticos, por ejemplo la función de densidad de la distribución normal tiene como expresión matemática la siguiente ecuación.

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{(y-\mu)}{\sigma}\right]^2}$$



Distribución Normal, mostrando los puntos de inflexión.

Donde se puede ver que la distribución queda totalmente representada por dos parámetros: μ (la media) y σ (la desviación estándar). Con las siguientes propiedades.

- Toda el área bajo la curva suma a 1.

- Los puntos de inflexión se localizan en $\mu - \sigma$ y $\mu + \sigma$.

- Entre $\mu - 4\sigma$ y $\mu + 4\sigma$ se encuentra la mayor parte del área bajo la curva (99.994%).

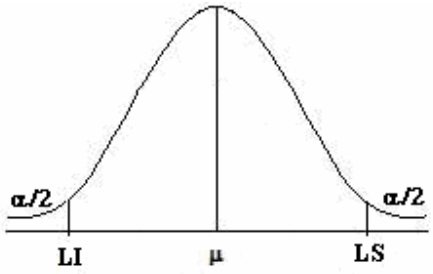
DISTRIBUCIÓN NORMAL ESTÁNDAR

- A una distribución normal con $\mu = 0$ y $\sigma = 1$ se le conoce como normal estándar y se representa por la variable *z* donde $z = \frac{(y-\mu)}{\sigma}$.

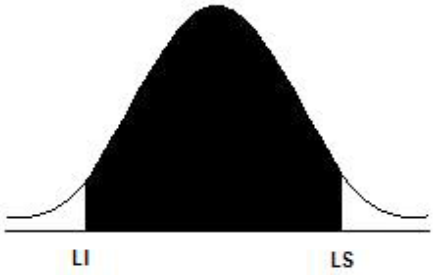
Su función densidad de probabilidad está dada por

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[z]^2}$$

Cada conjunto de datos genera una distribución con sus propios valores de μ, σ y f(y), además es difícil que el valor estimado a partir de la media sea exactamente μ, por lo que es común establecer intervalos de confianza en los que se espera que el verdadero valor se encuentre entre un límite inferior (LI) y uno superior (LS). Valores que al representarse en la distribución, como área bajo la curva, indican una probabilidad.



Distribución normal y Límites de confianza para la media



Área bajo la curva delimitada por los límites de confianza.

Los valores de Z asociados a LI y LS acotan o delimitan cierta proporción del área, de ahí la importancia de saber, por ejemplo, que -1.96 < Z < 1.96 delimita el 95% del área bajo la curva de una distribución normal y que el área que no está sombreada corresponde al complemento a 1, en este caso al 5%, que expresado en probabilidades se le conoce como nivel de significancia, α, y a (1 - α) como nivel de confianza.

De la misma forma el valor de -2.5756 < Z < 2.5756 delimita el 99%, con un complemento de 1% que dividido entre 2 corresponde al 0.5% ($\alpha_{(0.01)/2}$ = 0.005), lo interesante es que al asociar estos valores a los datos muestrales se pueden establecer intervalos de confianza para estimar los valores poblacionales.

TEOREMA CENTRAL DEL LÍMITE

Este teorema establece que la distribución de las medias muestrales es normal aún cuando las muestras se toman de una distribución no-normal.

Si $y_1, y_2, \ldots, y_n$ son resultados de una muestra de n observaciones independientes de una variable aleatoria Y con media μ y desviación σ, la media de las $\overline{Y}$´s se distribuirá aproximadamente en forma normal con media y varianza, respectivamente:

$$\mu_{\bar{y}} = \mu \qquad \text{y} \qquad \sigma^2_{\bar{y}} = \frac{\sigma^2}{n}$$

La aproximación es mucho mejor cuando n se hace grande. En general, la población de la cual se toman las

muestras no necesita ser normal, para que la distribución de las medias muestrales sea normal. Esto constituye lo más notorio y poderoso de este teorema.

ESTIMACIÓN (INTERVALOS DE CONFIANZA)

La estimación hace referencia al cálculo de intervalos de confianza para los parámetros de una distribución, a partir de datos muestrales.

Por ejemplo, para la estimación de la media se tiene:

$$P(LI < \mu < LS) = 1 - \alpha$$

que puede leerse como: la probabilidad de que el verdadero valor de μ esté en el intervalo acotado por LI y LS es 1-α, cuyo resultado numérico es $LI \leq \mu \leq LS$.

De aquí se pueden empezar a plantear las siguientes fórmulas de cálculo.

Cuadro 1: Intervalos de confianza para un parámetro.

Parámetro	Intervalo
1) μ Con varianza conocida o **n > 30** (donde n es el tamaño de muestra).	$\bar{y} \mp z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ $\bar{y} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
2) μ Con varianza desconocida o **n** $\leq$ **30**.	$\bar{y} \pm t_{1-\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$ $\bar{y} - t_{1-\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{1-\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$
3) π Proporción binomial con n grande	$p \mp z_{1-\frac{\alpha}{2}} \sqrt{\frac{pq}{n}}$ $p - z_{1-\frac{\alpha}{2}} \sqrt{\frac{pq}{n}} < \pi < p + z_{1-\frac{\alpha}{2}} \sqrt{\frac{pq}{n}}$
4) σ^2 Varianza distribución normal.	$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}$

El cuadro muestra los intervalos para los parámetros de una distribución normal: la media y la varianza. En la fórmula 1 se establece que la varianza es conocida, esto se logra cuando se tiene un proceso o fenómeno bien estudiado y se tiene una buena estimación del valor de la varianza poblacional. Cuando el tamaño de muestra es mayor a 30 se asume que $s^2 = \sigma^2$. En la fórmula 2 sólo se conoce la varianza muestral, así que para trabajar con ella hay que apoyarse en una distribución conocida como t de student, la cual también es simétrica y considera el manejo de n –1 grados de libertad. La fórmula 4 corresponde al intervalo para una varianza poblacional, a partir de la varianza muestral, aquí se debe utilizar una distribución conocida como Ji-cuadrada. Se requieren dos valores uno para el límite inferior y otro para el límite superior, ya que esta distribución no es simétrica y no tiene valores negativos, ya que al elevar al cuadrado un valor y luego sumarlo no hay posibilidades de obtener valores negativos.

Cuadro 2: Intervalos de confianza para dos parámetros.

5) $\mu_1 - \mu_2$ Con varianzas conocidas o $n_1 > 30$ y $n_2 > 30$.

$$(\bar{y}_1 - \bar{y}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{y}_1 - \bar{y}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

6) $\mu_1 - \mu_2$ Con varianzas desconocidas e iguales.

$$(\bar{y}_1 - \bar{y}_2) - t_{1-\frac{\alpha}{2},n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{y}_1 - \bar{y}_2) + t_{1-\frac{\alpha}{2},n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

con

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Es importante recordar que se asume o supone $\sigma_1^2 = \sigma_2^2$ y que g.l. = n_1+n_2-2

7) $\mu_1 - \mu_2$ Con varianzas desconocidas y diferentes.

$$(\bar{y}_1 - \bar{y}_2) - t_{1-\frac{\alpha}{2},\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{y}_1 - \bar{y}_2) + t_{1-\frac{\alpha}{2},\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

con

$$\nu = \text{grados de libertad} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Se asume o supone $\sigma_1^2 \neq \sigma_2^2$

8) Razón o cociente de varianzas de dos poblaciones normales

$$\frac{s_1^2}{s_2^2 (F_{1-\frac{\alpha}{2},n_1-1,n_2-1})} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2 (F_{\frac{\alpha}{2},n_1-1,n_2-1})}$$

Es importante notar los dos valores de F, aunque si se obtiene uno el otro es su inverso, cambiando los grados de libertad, esto es: $F_{\alpha,u,v} = \frac{1}{F_{1-\alpha,v,u}}$

Los cuadros 1 y 2 describen las fórmulas de cálculo para obtener intervalos de confianza para los parámetros de una población, considerando los valores muestrales (en este caso de la media y la varianza). Se debe aclarar que cada vez se utilizan menos, ya que el manejo numérico lo realiza el software, pero si es importante tenerlas en mente.

CONTRASTES O PRUEBAS DE HIPÓTESIS

Una hipótesis estadística es una aseveración acerca de los parámetros de una distribución de probabilidad.

Los procedimientos estadísticos de prueba de hipótesis se pueden utilizar para revisar la conformidad de los parámetros del proceso a sus valores especificados o para apoyar la modificación del proceso y lograr que se obtengan los valores deseados o especificados.

Para probar una hipótesis se toma una muestra aleatoria de la población en estudio, se calcula un estadístico de contraste adecuado, y se toma **la decisión de rechazar o no rechazar la hipótesis nula Ho.**

Ho: Hipótesis nula
Ha: Hipótesis alterna

Al realizar un contraste de hipótesis se pueden cometer dos tipos de errores

α = P{error tipo I}
α = P{rechazar Ho | Ho es verdadera}
β = P{error tipo II}
β = P{no rechazar Ho | Ho es falsa}

Para α = 0.05

Es importante notar que mientras más pequeño sea el valor de los extremos o colas de la distribución, se está más lejos de la zona de no rechazo de Ho.

Actualmente se hace referencia a $\hat{\alpha}$ (alfa gorro) o el nivel de significancia estimado a partir de los datos, también identificado como P-value o Significancia. Este valor indica si la probabilidad del error tipo I es mayor o menor que el nivel preestablecido (0.05 o 0.01) y que con el uso del software estadístico se ha vuelto fundamental para la interpretación de resultados. Reemplazando al uso de valores de tablas.

CONTRASTES DE HIPÓTESIS BASADOS EN LA DISTRIBUCIÓN NORMAL

VARIANZA(S) CONOCIDA(S)

9) Comparación de una media contra un valor definido por el investigador

Estadístico de contraste

$$z_c = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu = \mu_0$ Ha: $\mu \neq \mu_0$		$\lvert z_c \rvert > z_{1-\alpha/2}$
Ho: $\mu \geq \mu_0$ Ha: $\mu < \mu_0$		$z_c < -z_\alpha$
Ho: $\mu \leq \mu_0$ Ha: $\mu > \mu_0$		$z_c > z_\alpha$

10) Comparación de un par de medias

Estadístico de contraste

$$z_c = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$ Ha: $\mu_1 \neq \mu_2$ o $\mu_1 - \mu_2 \neq 0$		$\lvert z_c \rvert > z_{1-\alpha/2}$
Ho: $\mu_1 \geq \mu_2$ o $\mu_1 - \mu_2 \geq 0$ Ha: $\mu_1 < \mu_2$ o $\mu_1 - \mu_2 < 0$		$z_c < -z_\alpha$
Ho: $\mu_1 \leq \mu_2$ o $\mu_1 - \mu_2 \leq 0$ Ha: $\mu_1 > \mu_2$ o $\mu_1 - \mu_2 > 0$		$z_c > z_\alpha$

Al igual que en los intervalos de confianza, se supone varianza poblacional conocida, esto se logra cuando se tiene una buena estimación del valor de la varianza poblacional. Además, cuando el tamaño de muestra es mayor a 30 es común suponer que $s^2 = \sigma^2$.

VARIANZA(S) DESCONOCIDA(S)

11) Comparación de una media contra un valor definido por el investigador

Estadístico de contraste

$$t_c = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu = \mu_0$		$\lvert t_c \rvert > t_{1-\alpha/2,n-1}$
Ha: $\mu \neq \mu_0$		
Ho: $\mu \geq \mu_0$		$t_c < -t_{\alpha,n-1}$
Ha: $\mu < \mu_0$		
Ho: $\mu \leq \mu_0$		$t_c > t_{\alpha,n-1}$
Ha: $\mu > \mu_0$		

12) Comparación de un par de medias con varianzas desconocidas no diferentes

Estadístico de contraste

$$t_c = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$		$\lvert t_c \rvert > t_{1-\alpha/2,\nu}$
Ha: $\mu_1 \neq \mu_2$ o $\mu_1 - \mu_2 \neq 0$		
Ho: $\mu_1 \geq \mu_2$ o $\mu_1 - \mu_2 \geq 0$		$t_c < -t_{\alpha,\nu}$
Ha: $\mu_1 < \mu_2$ o $\mu_1 - \mu_2 < 0$		
Ho: $\mu_1 \leq \mu_2$ o $\mu_1 - \mu_2 \leq 0$		$t_c > t_{\alpha,\nu}$
Ha: $\mu_1 > \mu_2$ o $\mu_1 - \mu_2 > 0$		

Esta prueba corresponde a la comparación de dos medias, cuando las varianzas son iguales, en cuyo caso

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

con grados de libertad $\nu = n_1 + n_2 - 2$

13) Comparación de un par de medias, con varianzas desconocidas y diferentes

Estadístico de contraste

$$t_c = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$		$\lvert t_c \rvert > t_{1-\alpha/2,\nu}$
Ha: $\mu_1 \neq \mu_2$ o $\mu_1 - \mu_2 \neq 0$		
Ho: $\mu_1 \geq \mu_2$ o $\mu_1 - \mu_2 \geq 0$		$t_c < -t_{\alpha,\nu}$
Ha: $\mu_1 < \mu_2$ o $\mu_1 - \mu_2 < 0$		
Ho: $\mu_1 \leq \mu_2$ o $\mu_1 - \mu_2 \leq 0$		$t_c > t_{\alpha,\nu}$
Ha: $\mu_1 > \mu_2$ o $\mu_1 - \mu_2 > 0$		

Los grados de libertad se obtienen con

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Los contrastes 10 a 12 corresponden a pruebas t de student, ideal para muestras pequeñas, n menor que 30, o cuando no se tiene un buen estimador de la varianza poblacional y sólo se cuenta con la varianza muestral. Se debe recalcar que **para comparar dos medias, se debe realizar previamente un análisis de comparación de dos varianzas**.

COMPARACIÓN DE DATOS PAREADOS

El análisis de datos pareados es útil para comparar datos de antes y después, sobre todo donde es difícil conseguir material experimental en condiciones iniciales homogéneas. Por ejemplo, en investigaciones médicas, ya sea con seres humanos o animales de laboratorio, donde la diferencia entre el estado final y las condiciones iniciales antes de un tratamiento es mejor medida que sólo la medición final.

14) COMPARACIÓN DE DATOS PAREADOS

Estadístico de contraste

$$t_c = \frac{\overline{d} - \Delta_0}{\frac{s_d}{\sqrt{n}}}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\mu_d = \Delta_0$		$\lvert t_c \rvert > t_{1-\alpha/2,n-1}$
Ha: $\mu_d \neq \Delta_0$		
Ho: $\mu_d \leq \Delta_0$		$t_c > t_{\alpha,n-1}$
Ha: $\mu_d > \Delta_0$		
Ho: $\mu_d \geq \Delta_0$		$t_c < -t_{\alpha,n-1}$
Ha: $\mu_d < \Delta_0$		

15) COMPARACIÓN DE UNA PROPORCIÓN CONTRA UN VALOR SUPUESTO

Estadístico de contraste

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\pi = \pi_0$		$\lvert z_c \rvert > z_{1-\alpha/2}$
Ha: $\pi \neq \pi_0$		
Ho: $\pi \geq \pi_0$		$z_c < -z_\alpha$
Ha: $\pi < \pi_0$		
Ho: $\pi \leq \pi_0$		$z_c > z_\alpha$
Ha: $\pi > \pi_0$		

PRUEBAS PARA VARIANZAS

16) COMPARACIÓN DE UNA VARIANZA CONTRA UN VALOR SUPUESTO

Estadístico de contraste

$$\chi_c^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\sigma^2 = \sigma_0^2$		$\chi_c^2 > \chi^2_{1-\alpha/2,n-1}$
Ha: $\sigma^2 \neq \sigma_0^2$		$\chi_c^2 < \chi^2_{\alpha/2,n-1}$
Ho: $\sigma^2 \geq \sigma_0^2$		$\chi_c^2 < \chi^2_{\alpha,n-1}$
Ha: $\sigma^2 < \sigma_0^2$		
Ho: $\sigma^2 \leq \sigma_0^2$		$\chi_c^2 > \chi^2_{1-\alpha,n-1}$
Ha: $\sigma^2 > \sigma_0^2$		

17) COMPARACIÓN DE UN PAR DE VARIANZAS

Estadístico de contraste

$$F_c = \frac{s_1^2}{s_2^2}$$

Hipótesis	Estadístico de prueba	Regla de decisión Rechazar Ho si se cumple
Ho: $\sigma_1^2 = \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} = 1$		$F_c > F_{1-\alpha/2,n_1-1,n_2-1}$
Ha: $\sigma_1^2 \neq \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$		$F_c < F_{\alpha/2,n_1-1,n_2-1}$
Ho: $\sigma_1^2 \geq \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} \geq 1$		$F_c < F_{\alpha,n_1-1,n_2-1}$
Ha: $\sigma_1^2 < \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} < 1$		
Ho: $\sigma_1^2 \leq \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} \leq 1$		$F_c > F_{1-\alpha,n_1-1,n_2-1}$
Ha: $\sigma_1^2 > \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} > 1$		

Después de todo esto: ¿Cómo determinar intervalos de confianza y contrastes de hipótesis utilizando Statgraphics? Esto se muestra mediante ejemplos.

Contrastes con un Parámetro

Ejemplo 1

Se realizaron seis determinaciones del contenido de hidrógeno de un compuesto cuya composición teórica es del 9.55% en promedio, ¿Difiere el valor promedio del teórico?

%H 9.17, 9.09, 9.14, 9.10, 9.13, 9.27

Solución

El primer paso consiste en establecer el par de hipótesis, en otras palabras: quién es Ho y quién es Ha.

No realizar cálculo alguno si no sabe sabe contra que hipótesis se está trabajando

En este caso se tiene

Ho: $\mu = 9.55$ **Ha:** $\mu \neq 9.55$

La secuencia a seguir es:
0. Generar un archivo de datos con una sola columna, llamada porcentaje.
1. Seguir la secuencia

Describe -> Numeric Data -> One-Variable Analysis

2. Seleccionar la única variable y colocarla en la caja Data.

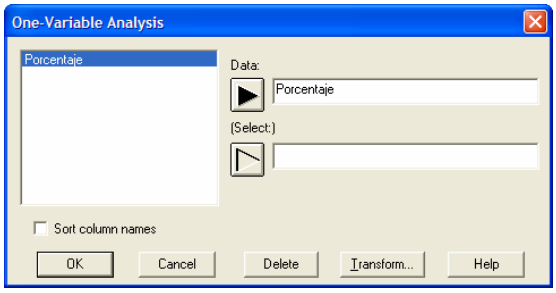


Figura 40. Diálogo para seleccionar variables.

3. En las opciones tabulares, asegurarse de seleccionar: Intervalos de Confianza y Prueba de Hipótesis.

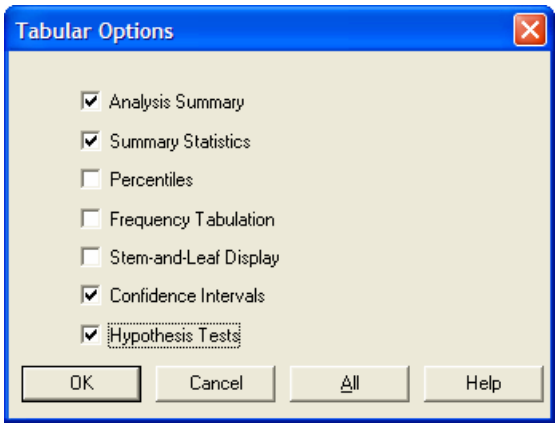


Figura 41. Opciones Tabulares.

4. En las opciones gráficas seleccionar Box and Wishker Plot.

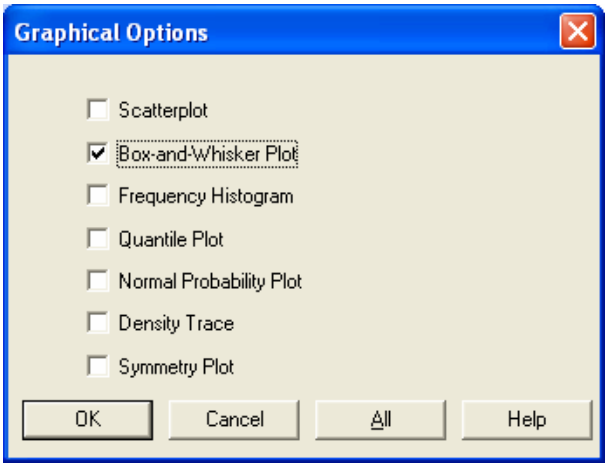


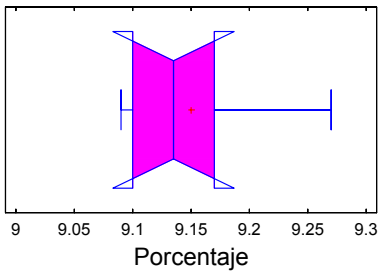
Figura 42. Opciones Gráficas.

Resultados

One-Variable Analysis - Porcentaje

```
Analysis Summary
Data variable: Porcentaje

6 values ranging from 9.09 to 9.27

The StatAdvisor
---------------
   This procedure is designed to summarize a
single sample of data.
It will calculate various statistics and graphs.
Also included in the procedure are confidence
intervals and hypothesis tests.  Use the Tabular
Options and Graphical Options buttons on the
analysis toolbar to access these different
procedures.
```

Box-and-Whisker Plot




```
Summary Statistics for Porcentaje

Count = 6
Average = 9.15
Variance = 0.00428
Standard deviation = 0.0654217
Minimum = 9.09
Maximum = 9.27
Range = 0.18
Stnd. skewness = 1.48497
Stnd. kurtosis = 1.21026
```

```
The StatAdvisor
---------------
   This table shows summary statistics for
Porcentaje. It includes measures of central
tendency, measures of variability, and measures
of shape. Of particular interest here are the
standardized skewness and standardized kurtosis,
which can be used to determine whether the
sample comes from a normal distribution. Values
of these statistics outside the range of -2 to
+2 indicate significant departures from
normality, which would tend to invalidate any
statistical test regarding the standard
deviation. In this case, the standardized
skewness value is within the range expected for
data from a normal distribution. The
standardized kurtosis value is within the range
expected for data from a normal distribution.
```

```
Confidence Intervals for Porcentaje
-----------------------------------
95.0% confidence interval for mean:
      9.15 +/- 0.0686561 [9.08134,9.21866]

95.0%   confidence   interval   for   Standard
deviation: [0.0408368 , 0.160454]
```

```
The StatAdvisor
---------------
   This pane displays 95.0% confidence intervals
for the mean and standard deviation of
Porcentaje. The classical interpretation of
these intervals is that, in repeated sampling,
these intervals Hill contain the true mean or
standard deviation of the population from which
the data come 95.0% of the time. In practical
terms, we can state with 95.0% confidence that
the true mean Porcentaje is somewhere between
9.08134 and 9.21866, while the true standard
deviation is somewhere between 0.0408368 and
0.160454.
   Both intervals assume that the population
from which the simple comes can be represented
by a normal distribution. While the confidence
interval for the mean is quite robust and not
very sensitive to violations of this assumption,
the confidence interval for the standard
deviation is quite sensitive. If the data do not
come from a normal distribution, the interval
for the Standard deviation may be incorrect. To
check whether the data come from a normal
distribution, select Summary Statistics from the
list of Tabular Options, or choose Normal
Probability Plot from the list of Graphical
Options.

Hypothesis Tests for Porcentaje

Sample mean = 9.15
Sample median = 9.135
```

```
t-test
------
Null hypothesis: mean = 0.0
Alternative: not equal

Computed t statistic = 342.59
P-Value = 4.02167E-12

Reject the null hypothesis for alpha = 0.05.
```

```
sign test
---------
Null hypothesis: median = 0.0
Alternative: not equal

Number of values below hypothesized median: 0
Number of values above hypothesized median: 6

Large sample test statistic = 2.04124
(continuity correction applied)
P-Value = 0.0412266

Reject the null hypothesis for alpha = 0.05.
```

```
signed rank test
----------------
Null hypothesis: median = 0.0
Alternative: not equal

Average rank of values below hypothesized
median: 0.0
Average rank of values above hypothesized
median: 3.5

Large sample test statistic = 2.09657
(continuity correction applied)
P-Value = 0.0360315

Reject the null hypothesis for alpha = 0.05.
```

```
The StatAdvisor
---------------
   This pane displays the results of three tests
concerning the center of the population from
which the sample of Porcentaje comes. The first
test is a t-test of the null hypothesis that the
mean   Porcentaje   equals   0.0   versus   the
alternative hypothesis that the mean Porcentaje
is not equal to 0.0. Since the P-value for this
test is less than 0.05, we can reject the null
hypothesis at the 95.0% confidence level.
The second test is a sign test of the null
hypothesis that the median Porcentaje equals 0.0
versus the alternative hypothesis that the
median Porcentaje is not equal to 0.0. It is
based on counting the number of values above and
below the hypothesized median. Since the P-value
for this test is less than 0.05, we can reject
the null hypothesis at the 95.0% confidence
level. The third test is a signed rank test of
the null hypothesis that the median Porcentaje
equals 0.0 versus the alternative hypothesis
that the median Porcentaje is not equal to 0.0.
It is based on comparing the average ranks of
values above and below the hypothesized median.
Since the P-value for this test is less than
0.05, we can reject the null hypothesis at the
95.0% confidence level. The sign and signed rank
tests are less sensitive to the presence of
outliers but are somewhat less powerful than the
t-test if the data all come from a single normal
distribution.
```

INTERPRETACIÓN

```
95.0% confidence interval for mean:
     9.15 +/- 0.0686561   [9.08134,9.21866]
```

Se tiene evidencia para afirmar con un 95% de confianza que el porcentaje promedio está entre los valores de 9.081 y 9.21, lo cual a su vez muestra que el porcentaje

promedio de 9.55, está muy por arriba de los límites de este intervalo.

Note que también se da el intervalo de confianza para la desviación estándar.

```
95.0%  confidence  interval  for  Standard
deviation: [0.0408368 , 0.160454]
```

La desviación estándar se encuentra entre 0.04 y 0.16 con una confianza de 95%

```
t-test
------
Null hypothesis: mean = 0.0
Alternative: not equal

Computed t statistic = 342.59
P-Value = 4.02167E-12

Reject the null hypothesis for alpha = 0.05.
```

Esta parte es interesante ya que Ho está dada en referencia al valor de 9.55, mientras que Statgraphhics, por regla general lo hace con respecto al valor cero, así que estos resultados no son muy útiles, por lo tanto, es necesario cambiar el valor de referencia.
Para esto, sobre la ventana de resultados de prueba de hipótesis, dar un clic derecho y seleccionar **Pane Options**, en la caja de diálogo que se despliega ahora si definir el valor de la media y si la prueba es bilateral o unilateral.

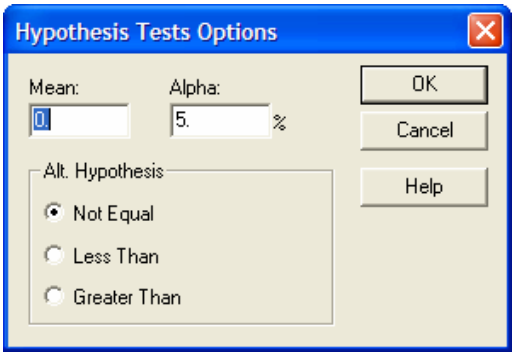


Figura 43. Pane Options.

Con el valor de 9.55 en la opción **Mean** se tienen los siguientes resultados.

```
Hypothesis Tests for Porcentaje

Sample mean = 9.15
Sample median = 9.135

t-test
------
Null hypothesis: mean = 9.55
Alternative: not equal

Computed t statistic = -14.9766
P-Value = 0.0000240279

Reject the null hypothesis for alpha = 0.05.
```

INTERPRETACIÓN

Como se vio en los intervalos de confianza, hay fuerte evidencia para afirmar con un 5% de significación (inclusive a un 1%) que el promedio es diferente de 9.55. Ya que el valor de P-value es 0.000024, muchísimo menor de 0.05.

NOTA: Se puede divertir probando que pasa cuando se selecciona la hipótesis alternativa con **Less Than** y con **Greater Than**.

Ejemplo 2

En un estudio sobre la utilización de agua en una ciudad pequeña se toma una muestra de 50 casas. La variable de interés, Y, es el número de galones de agua utilizado por día. Uno de los días, aleatoriamente elegido, de la semana se obtuvieron los siguientes valores.
El depósito de la ciudad es lo suficientemente grande para permitir una utilización media de 160 galones por día. ¿Da la impresión de que exista un problema de escasez de agua en la ciudad?

Datos

157	187	186	118	150
150	185	178	143	158
175	190	189	137	157
167	192	184	149	175
180	200	181	138	180
177	198	193	200	187
172	145	192	191	195
183	154	172	197	181
187	169	168	176	193
176	196	184	179	210

Ho: $\mu \leq 160$ **(no habrá escasez)**

Ha: $\mu > 160$ **(habrá escasez)**

SOLUCIÓN

0. Datos en un archivo con una sola columna, llamada galones de agua y 50 datos.

1. Seguir la secuencia
Describe -> Numeric Data -> One-Variable Analysis

2. Seleccionar la única variable y colocarla en la caja Data y dar OK.

3. En las opciones tabulares, asegurarse de seleccionar: Intervalos de Confianza y Prueba de Hipótesis.

Resultados

Antes de llegar a los resultados definitivos conviene que en la ventana de los intervalos y de las pruebas de hipótesis, entrar a **Pane Options** (con un clic derecho sobre la ventana) y probar los resultados con una prueba bilateral, o unilateral ya sea superior o inferior.

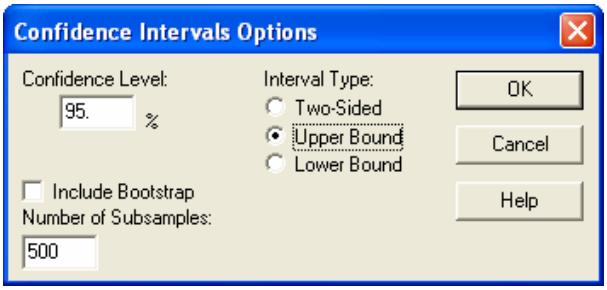
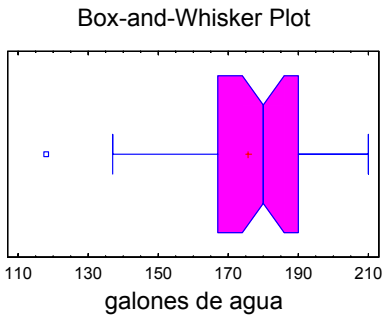


Figura 44. Pane Options para intervalos de confianza.

One-Variable Analysis - galones de agua

```
Analysis Summary

Data variable: galones de agua

50 values ranging from 118.0 to 210.0

The StatAdvisor
---------------
   This procedure is designed to summarize a
single sample of data.
It will calculate various statistics and graphs.
Also included in the procedure are confidence
intervals and hypothesis tests.  Use the Tabular
Options and Graphical Options buttons on the
analysis toolbar to access these different
procedures.
```




```
Summary Statistics for galones de agua

Count = 50
Average = 175.62
Variance = 381.669
Standard deviation = 19.5364
Minimum = 118.0
Maximum = 210.0
Range = 92.0
Stnd. skewness = -2.50938
Stnd. kurtosis = 0.559726

The StatAdvisor
---------------
   This table shows summary statistics for
galones de agua.  It includes measures of
central tendency, measures of variability, and
measures of shape.  Of particular interest here
are the standardized skewness and standardized
kurtosis, which can be used to determine whether
the sample comes from a normal distribution.
Values of these statistics outside the range of
-2 to +2 indicate significant departures from
normality, which would tend to invalidate any
statistical test regarding the standard
deviation.  In this case, the standardized
skewness value is not within the range expected
for data from a normal distribution.  The
standardized kurtosis value is within the range
expected for data from a normal distribution.

Confidence Bounds for galones de agua
-------------------------------------
95.0% lower confidence bound for mean:
          175.62 - 4.63208   [170.988]
95.0% lower confidence bound for standard
deviation: [16.7903]

The StatAdvisor
---------------
   This pane displays 95.0% lower confidence
bounds for the mean and standard deviation of
galones de agua.  The classical interpretation
of these bounds is that, in repeated sampling,
these bounds will bound the true mean or
standard deviation of the population from which
the data come 95.0% of the time.  In practical
terms, we can state with 95.0% confidence that
the true mean galones de agua is greater than or
equal to 170.988, while the true standard
deviation is greater than or equal to 16.7903.
   Both bounds assume that the population from
which the sample comes can be represented by a
normal distribution.  While the confidence bound
for the mean is quite robust and not very
sensitive to violations of this assumption, the
confidence bound for the Standard deviation is
quite sensitive.  If the data do not come from a
normal distribution, the interval for the
standard deviation may be incorrect.  To check
whether the data come from a normal
distribution, select Summary Statistics from the
list of Tabular Options, or cose Normal
Probability Plot from the list of Graphical
Options.

Hypothesis Tests for galones de agua

Sample mean = 175.62
Sample median = 180.0

t-test
------
Null hypothesis: mean = 160.0
Alternative: greater than

Computed t statistic = 5.65357
P-Value = 5.59273E-7
```

```
Reject the null hypothesis for alpha = 0.05.

sign test
---------
Null hypothesis: median = 160.0
Alternative: greater than

Number of values below hypothesized median: 12
Number of values above hypothesized median: 38

Large sample test statistic = 3.53553
(continuity correction applied)
P-Value = 0.000203517

Reject the null hypothesis for alpha = 0.05.

signed rank test
----------------
Null hypothesis: median = 160.0
Alternative: greater than
Average rank of values below hypothesized
median: 14.3333
Average rank of values above hypothesized
median: 29.0263

Large sample test statistic = 4.4893 (continuity
correction applied)
P-Value = 0.00000357616

Reject the null hypothesis for alpha = 0.05.

The StatAdvisor
---------------
   This pane displays the results of three tests
concerning the center of the population from
which the sample of galones de agua comes.  The
first test is a t-test of the null hypothesis
that the mean galones de agua equals 160.0
versus the alternative hypothesis that the mean
galones de agua is greater than 160.0.  Since
the P-value for this test is less than 0.05, we
can reject the null hypothesis at the 95.0%
confidence level.  The second test is a sign
test of the null hypothesis that the median
galones de agua equals 160.0 versus the
alternative hypothesis that the median galones
de agua is greater than 160.0.  It is based on
counting the number of values above and below
the hypothesized median.  Since the P-value for
this test is less than 0.05, we can reject the
null hypothesis at the 95.0% confidence level.
The third test is a signed rank test of the null
hypothesis that the median galones de agua
equals 160.0 versus the alternative hipótesis
that the median galones de agua is greater than
160.0.  It is based on comparing the average
ranks of values above and below the hypothesized
median.  Since the P-value for this test is less
than 0.05, we can reject the null hypothesis at
the 95.0% confidence level.  The sign and signed
rank tests are less sensitive to the presence of
outliers but are somewhat less powerful than the
t-test if the data all come from a single normal
distribution.
```

INTERPRETACIÓN

Primero y antes que nada ¿por qué una prueba de t y no una de Z? Pues simple y sencillamente porque para n's mayores de 30 los valores de t y Z tienden a igualarse. Esto es más claro para n mayor que 200.

Entonces para qué eso de muestras grandes, muestras pequeñas, varianza conocida o varianza desconocida. Pues simple y sencillamente para saber que prueba es la adecuada y que estadístico de contraste es el más adecuado.

Luego, una hipótesis unilateral o bilateral, en términos reales no se pide que sea diferente de 160, ya que podría ser diferente de 160 por ser mayor o por ser menor. Lo que se pide es saber si el consumo es mayor que 160 galones, por lo que en **Pane Options** se pide el menor valor posible que consume esta muestra de valores, por lo tanto sería el **lower bound**.

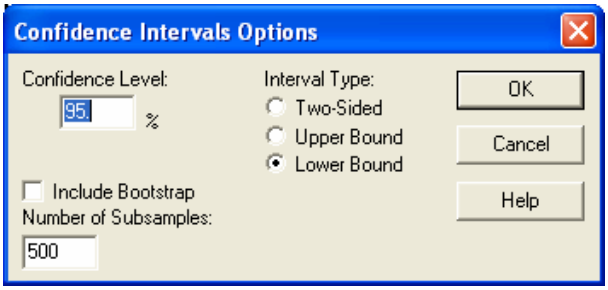


Figura 45. Pane Options para intervalos d confianza.

Ahora si:

1)

```
95.0% lower confidence bound for mean:
          175.62 - 4.63208   [170.988]
95.0% lower confidence bound for standard
deviation: [16.7903]
In practical terms, we can state with 95.0%
confidence that the true mean galones de agua is
greater than or equal to 170.988, while the true
standard deviation is greater than or equal to
16.7903.
```

Se tiene evidencia para afirmar con un 95% de confianza que el consumo promedio de agua por casa está por arriba de los 170.988 galones.

2)

```
Hypothesis Tests for galones de agua

Sample mean = 175.62
Sample median = 180.0


t-test
------
Null hypothesis: mean = 160.0
Alternative: greater than

Computed t statistic = 5.65357
P-Value = 5.59273E-7

Reject the null hypothesis for alpha = 0.05.
```

Al rechazar la hipótesis nula, se tiene evidencia estadística, al 5% de significación, para recomendar que disminuyan su consumo de agua si no quieren tener problemas de desabasto en un futuro cercano.

Ejemplo 3

Los estudiantes de tercero de primaria alcanzan una calificación media de 75 y una varianza de 50 en la prueba Robertson del conocimiento de acontecimientos actuales. Si se eligen aleatoriamente 5 niños de tercer grado de una escuela "abierta" quienes obtienen puntajes de 85, 92, 91, 91, 91. Bajo la hipótesis de que la escuela abierta fue elegida por los padres de los niños porque en ella se fomenta la agilidad mental, sus puntajes deben ser más homogéneos que los puntajes de la población general de los niños de tercer grado. Probar esta hipótesis.

Ho: $\sigma^2 \geq 50$ **(puntajes no más homogéneos)**
Ha: $\sigma^2 < 50$ **(puntajes menos variables o más homogéneos).**

SOLUCIÓN

Seguir la misma secuencia de solución que para los ejemplos 1 y 2, sólo hay que cuidar que en las opciones tabulares tengamos seleccionados Summary Statistics y Confidence intervals.

Resultados

```
One-Variable Analysis - calificaciones
Summary Statistics for calificaciones

Count = 5
Average = 90.0
Variance = 8.0
Standard deviation = 2.82843
Minimum = 85.0
Maximum = 92.0
Range = 7.0
Stnd. skewness = -1.91632
Stnd. kurtosis = 2.08962

The StatAdvisor
---------------
   This table shows summary statistics for
calificaciones.  It includes measures of central
tendency, measures of variability, and measures
of shape.  Of particular interest here are the
standardized skewness and standardized kurtosis,
which can be used to determine whether the
sample comes from a normal distribution.  Values
of these statistics outside the range of -2 to
+2 indicate significant departures from
normality, which would tend to invalidate any
statistical test regarding the standard
deviation.  In this case, the standardized
skewness value is within the range expected for
data from a normal distribution.  The
standardized kurtosis value is not within the
range expected for data from a normal
distribution.

Confidence Bounds for calificaciones
-----------------------------------
95.0% upper confidence bound for mean:
            90.0 + 2.6966   [92.6966]
95.0% upper confidence bound for standard
deviation: [6.71004]

The StatAdvisor
---------------
   This pane displays 95.0% upper confidence
bounds for the mean and standard deviation of
calificaciones. The classical interpretation of
these bounds is that, in repeated sampling,
these bounds will bound the true mean or
standard deviation of the population from which
the data come 95.0% of the time. In practical
terms, we can state with 95.0% confidence that
the true mean calificaciones is less than or
equal to 92.6966, while the true standard
deviation is less than or equal to 6.71004.
   Both bounds assume that the population from
which the sample comes can be represented by a
normal distribution.  While the confidence bound
for the mean is quite robust and not very
sensitive to violations of this assumption, the
confidence bound for the Standard deviation is
quite sensitive. If the data do not come from a
normal distribution, the interval for the
standard deviation may be incorrect. To check
whether the data come from a normal
distribution, select Summary Statistics from the
list of Tabular Options, or cose Normal
Probability Plot from the list of Graphical
Options.
```

INTERPRETACIÓN

Extrayendo algunos fragmentos del StatAdvisor se tiene:

```
while the true standard deviation is less than
or equal to 6.71004.
```

La desviación estándar es menor o igual a 6.71004, por lo tanto, la varianza es el cuadrado de 6.71004, o 45.0246, que está muy por debajo de 50.

Por consiguiente, se tiene evidencia al 5% de significación de que los puntajes de los alumnos de la escuela abierta son más homogéneos que los de los alumnos normales.

Ejemplo 4

Un fabricante de pilas de níquel-cadmio selecciona al azar 120 placas de níquel para celdas de prueba, les pone y les quita corriente una cierta cantidad de veces y determina que 15 de ellas se han ampollado. ¿Es esto una evidencia suficente para llegar al resultado de que

más del 10% de todas las placas se ampollan bajo estas condiciones de prueba? Use una significación del 1%

Solución:

El primer paso consiste en establecer el par de hipótesis, en otras palabras: quién es Ho y quién es Ha.

En este caso se tiene

Ho: $\pi \leq 0.10$ **Ha:** $\pi > 0.10$

La secuencia a seguir es:

1. Seguir la secuencia

Describe -> Numeric Data -> Hypothesis tests

2. Seleccione el parámetro con el que está trabajando (Proporción binomial), el valor supuesto en la hipótesis nula, el valor del estimador muestral ($\frac{15}{120} = 0.125$) y el tamaño de la muestra y luego presione OK.

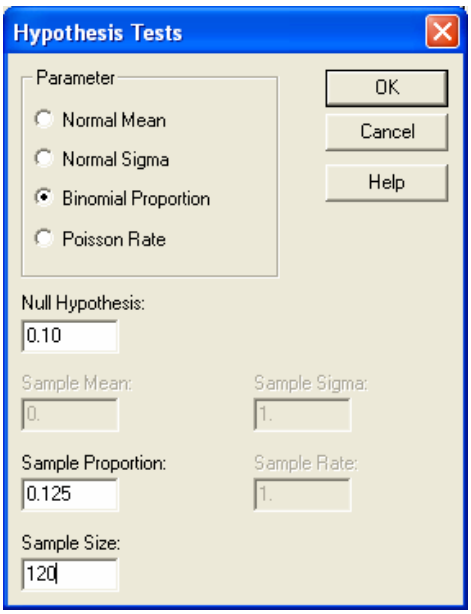


Figura 46. Diálogo para seleccionar el parámetro, la hipótesis, el estimador y el tamaño de la muestra.

3. Sobre la pantalla de resultados, presione el botón derecho para obtener las opciones de análisis de las pruebas de hipótesis, selecciones la hipótesis alterna **Greater Than** y el valor de Alpha (1%), presione OK.

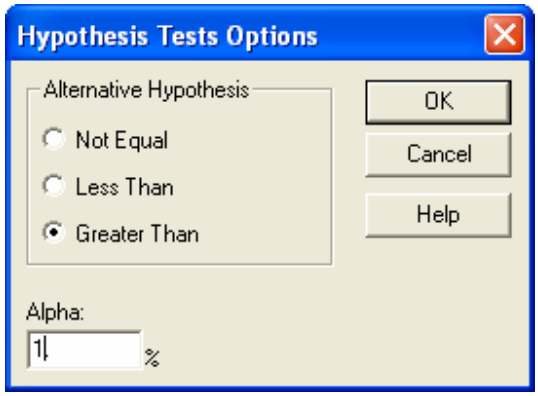


Figura 47. Opciones para contrastes de Hipótesis.

Resultados

```
Hypothesis Tests
----------------
Sample proportion = 0.125
Sample size = 120

Approximate 99.0% lower confidence bound for p:
[0.0640908]

Null Hypothesis: proportion = 0.1
Alternative: greater than
P-Value = 0.218163
Do not reject the null hypothesis for alpha =
0.01.

The StatAdvisor
---------------
   This analysis shows the results of performing
a hypothesis test
concerning the proportion (theta) of a binomial
distribution.  The two
hypotheses to be tested are:

   Null hypothesis:        theta = 0.1
   Alternative hypothesis: theta > 0.1

In this sample of 120 observations, the sample
proportion equals
0.125.  Since the P-value for the test is
greater than or equal to
0.01, the null hypothesis cannot be rejected at
the 99.0% confidence
level.  The confidence bound shows that the
values of theta supported
by the data are greater than or equal to
0.0640908.
```

INTERPRETACIÓN

Como el p-value es 0.218163 > 0.01, no se rechaza la hipótesis nula y se concluye que la proporción de placas que se ampollan no es mayor que el 10%.

Observe que en los ejemplos 1, 2 y 3 se tienen datos crudos y la secuencia de análisis es diferente a la utilizada en el ejemplo 4. Si en los ejemplos 1, 2 y 3 no se tuvieran datos sino un valor muestral calculado

por algún otro investigador y se quisiera contrastar contra el valor supuesto o especificado, la secuencia de análisis sería la indicada en el ejemplo 4.

Ejemplo 5

En un estudio publicado en el Journal of Occupational and Organizacional Psychology (Diciembre 1992) se investigó la relación entre el estatus en el empleo y la salud mental. A cada uno de los hombres de una muestra de 49 desempleados se le aplicó el test sobre salud mental General Health Questionnaire (GHQ), donde valores pequeños indica mejor salud mental. La media y desviación estándar de esta muestra fueron: $\bar{x} = 10.94$ y $s = 5.10$, respectivamente. Determine si la media en el GHQ para todos los desempleados excede 10. Enuncie y contraste el par de hipótesis correspodiente.
(Problema tomado de: McClave, J. T., et al, (1997), Statistics, 7th. Edition, Prentice Hall, Inc., U. S. A.

SOLUCIÓN

1. El primer paso consiste en establecer el par de hipótesis, en otras palabras: quién es Ho y quién es Ha.

En este caso se tiene

Ho: $\mu \leq 10$ **Ha:** $\mu > 10$

2. La secuencia a seguir es

Describe -> Numeric Data -> Hypothesis tests

Seleccionar el parámetro con el que está trabajando (Normal Mean), el valor supuesto en la hipótesis nula (10), los valores de los estimadores muestrales, de la media 10.94) y de la desviación estándar (5.10), el tamaño de la muestra (49) y luego presione OK.

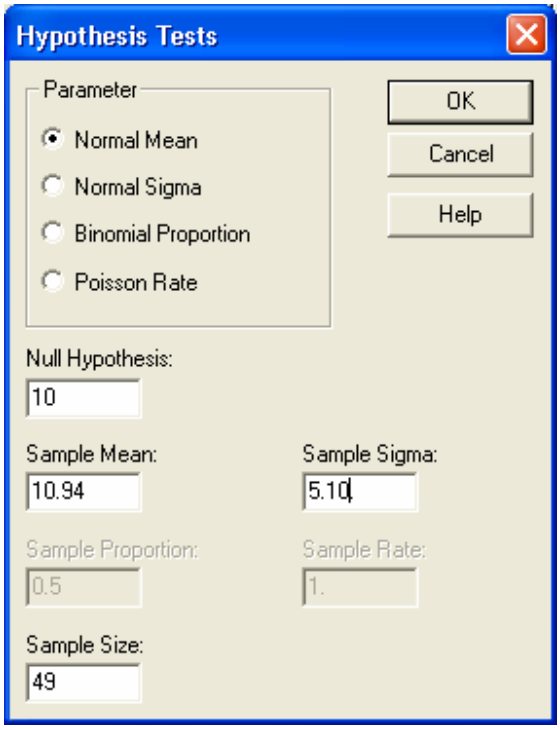


Figura 48. Diálogo para seleccionar el parámetro, la hipótesis, los estimadores de la media y desviación estándar y el tamaño de la muestra.

3. Sobre la pantalla de resultados, presione el botón derecho para obtener las opciones de análisis de las pruebas de hipótesis, selecciones la hipótesis alterna **Greater Than** y el valor de Alpha (5%), presione OK.

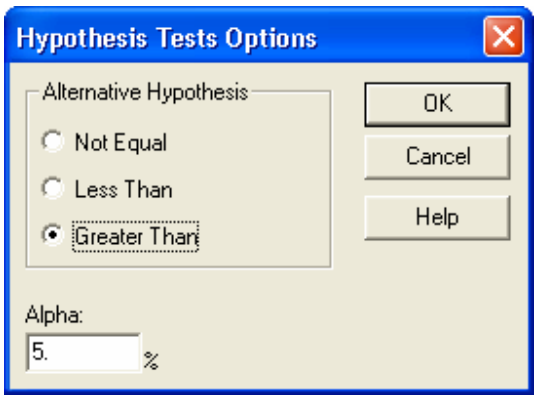


Figura 49. Opciones para contrastes de Hipótesis.

Resultados

```
Hypothesis Tests
----------------
Sample mean = 10.94
Sample standard deviation = 5.1
Sample size = 49

95.0% lower confidence bound for mean:
10.94 - 1.22198    [9.71802]
```

```
Null Hypothesis: mean = 10.0
Alternative: greater than
Computed t statistic = 1.2902
P-Value = 0.101582
Do not reject the null hypothesis for alpha =
0.05.

The StatAdvisor
---------------
   This analysis shows the results of performing
a hypothesis test
concerning the mean (mu) of a normal
distribution.  The two hypotheses
to be tested are:

   Null hypothesis:        mu = 10.0
   Alternative hypothesis: mu > 10.0

Given a sample of 49 observations with a mean of
10.94 and a standard
deviation of 5.1, the computed t statistic
equals 1.2902.  Since the
P-value for the test is greater than or equal to
0.05, the null
hypothesis cannot be rejected at the 95.0%
confidence level.  The
confidence bound shows that the values of mu
supported by the data are
greater than or equal to 9.71802.
```

INTERPRETACIÓN

Como el p-value es 0.101582 > 0.05, no se rechaza la hipótesis nula y se concluye que la media de la salud mental de los desempleados no es mayor que el 10.

Ejemplo 6

Usando los datos del ejemplo 4, enuncie y contraste la hipótesis de que la desviación estándar en el test GHQ e la salud mental de la población de desempleados es mayor que 4.5.

Solución

1. El primer paso consiste en establecer el par de hipótesis, en otras palabras: quién es Ho y quién es Ha.

En este caso se tiene

Ho: $\sigma \leq 4.5$ **Ha:** $\sigma > 4.5$

2. La secuencia a seguir es

Describe -> Numeric Data -> Hypothesis tests

Seleccione el parámetro con el que está trabajando (Normal Sigma), el valor supuesto en la hipótesis nula (10), los valores de los estimadores muestrales, de la media 10.94) y de la desviación estándar (5.10), el tamaño de la muestra (49) y luego presione OK.

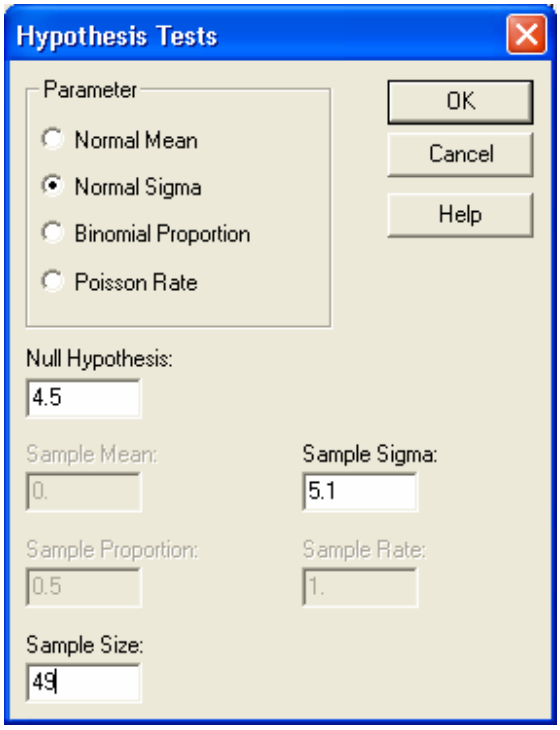


Figura 50. Diálogo para seleccionar el parámetro, la hipótesis, el estimador de la desviación estándar y el tamaño de la muestra.

3. Sobre la pantalla de resultados, presione el botón derecho para obtener las opciones de análisis de las pruebas de hipótesis, selecciones la hipótesis alterna **Greater Than** y el valor de Alpha (5%), presione OK.

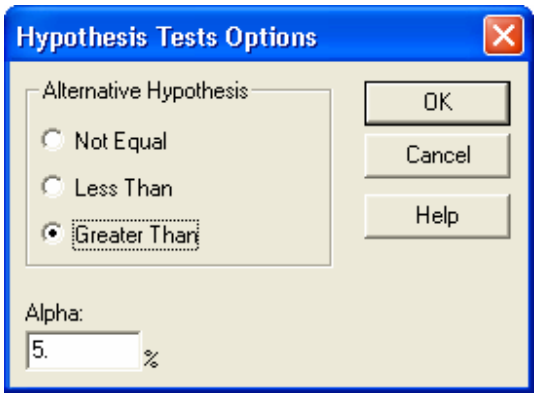


Figura 51. Opciones para contrastes de Hipótesis.

Resultados

```
Hypothesis Tests
----------------
Sample standard deviation = 5.1
Sample size = 49

95.0% lower confidence bound for sigma:
[4.37688]

Null Hypothesis: standard deviation = 4.5
Alternative: greater than
Computed chi-squared statistic = 61.6533
P-Value = 0.0890958
Do not reject the null hypothesis for alpha =
0.05.

The StatAdvisor
---------------
   This analysis shows the results of performing
a hypothesis test
concerning the standard deviation (sigma) of a
normal distribution.
The two hypotheses to be tested are:

   Null hypothesis:        sigma = 4.5
   Alternative hypothesis: sigma > 4.5

Given a sample of 49 observations with a
standard deviation of 5.1,
the computed chi-square statistic equals
61.6533.  Since the P-value
for the test is greater than or equal to 0.05,
the null hypothesis
cannot be rejected at the 95.0% confidence
level.  The confidence
bound shows that the values of sigma supported
by the data are greater
than or equal to 4.37688.
```

INTERPRETACIÓN

Como el p-value es 0.0890958 > 0.05, no se rechaza la hipótesis nula y se concluye que la desviación estándar de la salud mental de los desempleados no es mayor que el 4.5.

COMPARACIÓN DE UN PAR DE PARÁMETROS POBLACIONALES.

Ejemplo 7

Se analizó el contenido de silicio de una muestra de agua por dos métodos independientes, en un intento por mejorar la precisión de la determinación. De acuerdo a los siguientes datos.

Método original	Método modificado
149 ppm	150 ppm
139	147
135	152
140	151
155	145

¿Qué se puede decir de las respuestas promedio y de su precisión?

SOLUCIÓN

El primer paso consiste en establecer el par de hipótesis, en otras palabras: quién es Ho y quién es Ha.

Para las respuestas promedio se tiene:

Ho: $\mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$

Ha: $\mu_1 \neq \mu_2$ o $\mu_1 - \mu_2 \neq 0$

Para la precisión se tienen las hipótesis

Ho: $\sigma_1^2 = \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} = 1$ y Ha: $\sigma_1^2 \neq \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

Este par de hipótesis permite probar que las precisiones son diferentes.

Ho: $\sigma_1^2 \leq \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} \leq 1$ y Ha: $\sigma_1^2 > \sigma_2^2$ o $\frac{\sigma_1^2}{\sigma_2^2} > 1$

Este último par de hipótesis permite probar que el método modificado (2) tiene menos variabilidad (mayor precisión) que el método original (1).

Secuencia de análisis

0. Generar un archivo de datos con dos columnas, llamadas método y ppm, ambas numéricas.

1. Del menú seguir las opciones
Compare -> Two samples -> Two-sample comparison

2. En la caja de diálogo seleccionar la opción **Data and Code Columns** en **Input**. Colocar la variable **ppm en Data** y **método en Sample Code.** Aunque los datos también pueden "teclearse" en dos columnas.

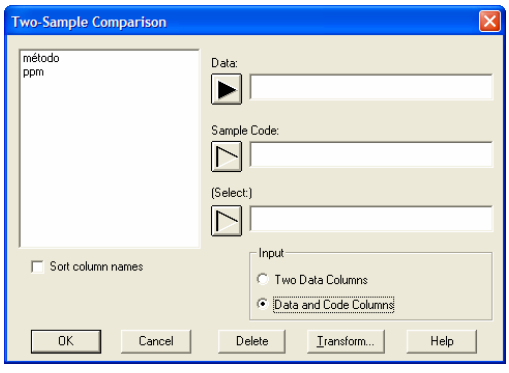


Figura 52. Data and Code Columns.

3. Dar un clic sobre el botón **OK** y listos para analizar los resultados.

4. Seleccionar las opciones tabulares

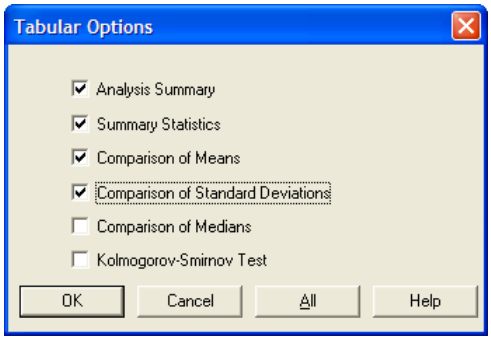


Figura 53. Opciones tabulares para comparación de medias.

Para obtener los resultados deseados se deben activar **Comparison of means** (comparación de medias) y **Comparison of Standar Deviations** (Comparación de desviaciones estándar).

RESULTADOS

Two-Sample Comparison - ppm & método

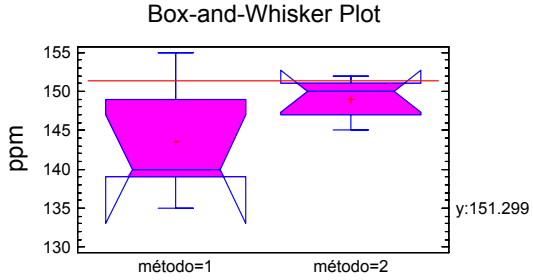
Analysis Summary for ppm

Sample 1: método=1
Sample 2: método=2

Sample 1: 5 values ranging from 135.0 to 155.0
Sample 2: 5 values ranging from 145.0 to 152.0

The StatAdvisor

This procedure is designed to compare two samples of data. It Will calculate various statistics and graphs for each sample, and it will run several tests to determine whether there are statistically significant differences between the two samples.



Summary Statistics for ppm

	método=1	método=2
Count	5	5
Average	143.6	149.0
Median	140.0	150.0
Mode		
Variance	66.8	8.5
Standard deviation	8.17313	2.91548
Minimum	135.0	145.0
Maximum	155.0	152.0
Range	20.0	7.0

The StatAdvisor

This table shows summary statistics for the two samples of data. Other tabular options within this analysis can be used to test whether differences between the statistics from the two samples are statistically significant. Of particular interest here are the standardized skewness and standardized kurtosis, which can be used to determine whether the samples come from normal distributions. Values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate the tests which compare the standard deviations. In this case, both standardized skewness values are within the range expected. Both standardized kurtosis values are within the range expected.

Comparison of Means for ppm

95.0% confidence interval for mean of método=1:
143.6 +/- 10.1483 [133.452,153.748]
95.0% confidence interval for mean of método=2:
149.0 +/- 3.62005 [145.38,152.62]
95.0% confidence interval for the difference between the means assuming equal variances:
-5.4 +/- 8.94898 [-14.349,3.54898]

t test to compare means
Null hypothesis: mean1 = mean2
Alt. hypothesis: mean1 NE mean2
assuming equal variances: t = -1.39149
P-value = 0.201547

The StatAdvisor

This option runs a t-test to compare the means of the two samples. It also constructs confidence intervals or bounds for each mean and for the difference between the means. Of particular interest is the confidence interval for the difference between the means, which extends from -14.349 to 3.54898. Since the interval contains the value 0.0, there is not a statistically significant difference between the means of the two samples at the 95.0% confidence level.

A t-test may also be used to test a specific hypothesis about the difference between the means of the populations from which the two samples come. In this case, the test has been constructed to determine whether the difference between the two means equals 0.0 versus the alternative hypothesis that the difference does not equal 0.0. Since the computed P-value is not less than 0.05, we cannot reject the null hypothesis.

NOTE: these results assume that the variances of the two simples are equal. In this case, that assumption appears to be reasonable based on the results of an F-test to compare the standard deviations. You can see the results of that test by selecting Comparison of Standard Deviations from the Tabular Options menu.

Comparison of Standard Deviations for ppm

	método=1	método=2
Standard deviation	8.17313	2.91548
Variance	66.8	8.5
Df	4	4

Ratio of Variances = 7.85882

95.0% Confidence Intervals
Standard deviation of método=1: [4.89679,23.4859]
Standard deviation of método=2: [1.74676,8.37778]
Ratio of Variances: [0.818242,75.4803]

F-test to compare Standard Deviations

Null hypothesis: sigma1 = sigma2
Alt. hypothesis: sigma1 NE sigma2
F = 7.85882 P-value = 0.0707003
The StatAdvisor

This option runs an F-test to compare the variances of the two samples. It also constructs confidence intervals or bounds for each standard deviation and for the ratio of the variances. Of particular interest is the confidence interval for the ratio of the variances, which extends from 0.818242 to 75.4803. Since the interval contains the value 1.0, there is not a statistically significant difference between the standard deviations of the two samples at the 95.0% confidence level.

An F-test may also be used to test a specific hypothesis about the standard deviations of the populations from which the two simples come. In this case, the test has been constructed to determine whether the ratio of the standard deviations equals 1.0 versus the alternative hypothesis that the ratio does not equal 1.0. Since the computed P-value is not less than 0.05, we cannot reject the null hypothesis.

IMPORTANT NOTE: the F-tests and confidence intervals shown here depend on the samples having come from normal distributions. To test this assumption, select Summary Statistics from the list of Tabular Options and check the standardized skewness and standardized kurtosis values.

INTERPRETACIÓN

1. En la parte se Summary Analysis es importante considerar la siguiente nota:

Of particular interest here are the standardized skewness and standardized kurtosis, which can be used to determine whether the samples come from normal distributions. Values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate the tests which compare the standard deviations. In this case, both standardized skewness values are within the range expected. Both standardized kurtosis values are within the range expected.

La mejor manera de ver si se cumple la normalidad consiste en dar un clic derecho en la **ventana de resultados del Summary** y solicitar **Pane Options.**

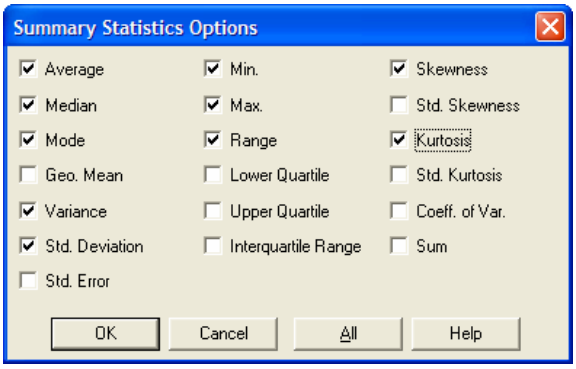


Figura 54. Pane options para Summary Statistics.

Seleccionar **Skewness** y **Kurtosis,** de las opciones, para asegurarse que no sale del rango que garantiza normalidad en los datos. **Cabe hacer notar que los valores de Skewness y de kurtosis son valores estandarizados.**

2. Comparando las medias se tiene

Null hypothesis: mean1 = mean2
Alt. hypothesis: mean1 NE mean2
assuming equal variances: t = -1.39149
P-value = 0.201547

Como P-value es mayor que 0.05 se tiene evidencia estadística para NO rechazar Ho, de manera que la respuesta promedio es semejante en ambos métodos, siempre y cuando las varianzas sean iguales.

NOTA: Siempre y cuando las varianzas sean iguales. Esto implica que primero se debe probar la igualdad de varianzas.

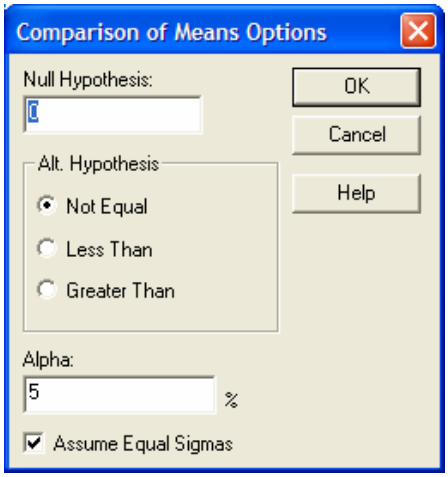


Figura 55. Pane Options para la Comparación de medias.

3. Comparando las varianzas se tiene

```
     Ratio of Variances: [0.818242,75.4803]

F-test to Compare Standard Deviations

   Null hypothesis: sigma1 = sigma2
   Alt. hypothesis: sigma1 NE sigma2
   F = 7.85882    P-value = 0.0707003
```

Como el intervalo para el cociente de varianzas va de 0.818 a 75.48, incluye el 1, significa que las varianzas no son significativamente diferentes. Además en la prueba de desviaciones estándar se tiene un P-value de 0.707003 > 0.05, por lo que se puede asumir que las dos varianzas (desviaciones estándar) no son diferentes, lo que sustenta la validez de las conclusiones en la comparación de medias, ya que si las varianzas fueran diferentes se tiene que activar el **Pane Options** de la comparación de medias y **desactivar la opción de suponer desviaciones estándar iguales**.

5. En este caso la igualdad de desviaciones se "ve" medio rara, ya que el cociente de varianzas está muy lejos del 1.

```
-----------------------------------------------
                    método=1          método=2
-----------------------------------------------
Standard deviation  8.17313           2.91548
Variance            66.8              8.5
Df                  4                 4

     Ratio of Variances = 7.85882
```

6. Realizando la comparación de varianzas para una prueba unilateral (de una cola), a la cual se llega mediante el **Pane Options de la comparación de desviaciones.**

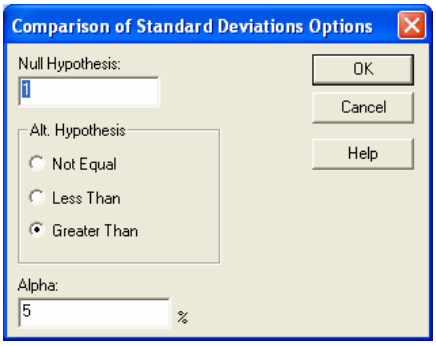


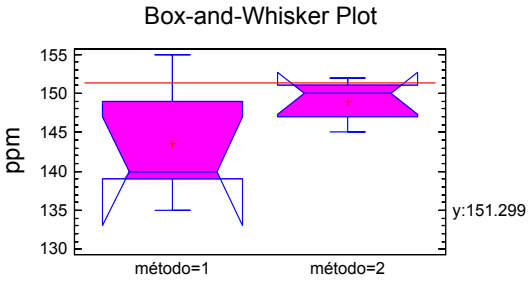
Figura 56. Pane Options para la
Comparación de desviaciones estándar.

Seleccionando la Hipótesis alternativa como **Greater Than.**

```
F-test to Compare Standard Deviations
   Null hypothesis: sigma1 = sigma2
   Alt. hypothesis: sigma1 > sigma2

   F = 7.85882    P-value = 0.0353502
```

Notar que ahora si se tiene evidencia estadística para afirmar con un 5% de soignificación que el método modificado (método 2) tiene más precisión que el método original, ya que se rechaza la hipótesis nula de que la varianza del método 1 no es mayor que la varianza del método 2.

7. Un comparativo visual de este análisis se tiene en el Gráfico de cajas y alambres.



Nótese la diferencia de variabilidad por la altura de las cajas.

Ejemplo 8

Se realizó un estudio para probar que un programa de ejercicios regulares moderadamente activos beneficia a pacientes que previamente han sufrido un infarto al miocardio. Once individuos participan en el estudio, de manera que antes de iniciar el programa se les determina la capacidad de trabajo midiendo el tiempo que tardan en alcanzar una tasa de 160 latidos por minuto mientras caminaban sobre una banda sin fin. Después de 25

semanas de ejercicio controlado, se repitieron las medidas, encontrando los siguientes resultados.

Sujeto	1	2	3	4	5
Antes	7.6	9.9	8.6	9.5	8.4
Después	14.7	14.1	11.8	16.1	14.7

Sujeto	6	7	8	9	10	11
Antes	9.2	6.4	9.9	8.7	10.3	8.3
Después	14.1	13.2	14.9	12.2	13.4	14.0

¿Realmente funciona el programa de ejercicios?

SOLUCIÓN

El primer paso consiste en establecer el par de hipótesis, en otras palabras: quién es Ho y quién es Ha.

Para las respuestas promedio se tiene:

Ho: $\mu_d \geq 0$ **o** $\mu_A - \mu_B \geq 0$

Ha: $\mu_d < 0$ **o** $\mu_A - \mu_B < 0$

Este par de hipótesis permite probar si hay diferencias entre los resultados de antes y los de después, ya que las mediciones se hacen sobre el mismo individuo.

Secuencia de análisis

0. Datos en un archivo con dos columnas, una para los datos de antes y una para los datos de después. También se puede agregar una columna para los sujetos o individuos, pero esta no se utiliza en el análisis.

1. Del menú seguir

Compare -> Two samples -> Paired-sample comparison

2. Colocar las variables antes y después en Sample 1 y Sample 2 de la caja de diálogo que se despliega.

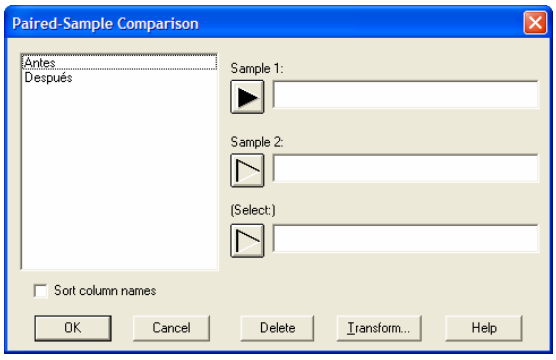


Figura 57. Comparación de medias pareadas.

3. Dar un clic sobre el botón OK y listos para empezar a ver los resultados.

4. En la ventana de resultados, dar un clic sobre el icono de opciones tabulares. Asegurando que esté activo Confidence Intervals e Hypothesis Test

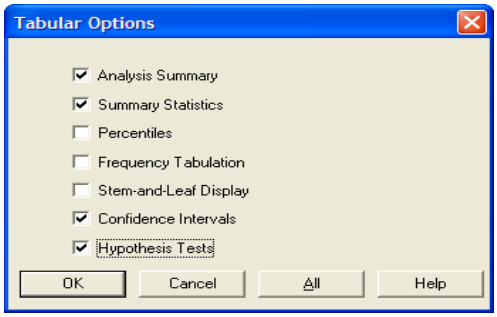


Figura 58. Tabular Options.

5. En las opciones gráficas seleccionar únicamente Box and Whisker Plot.

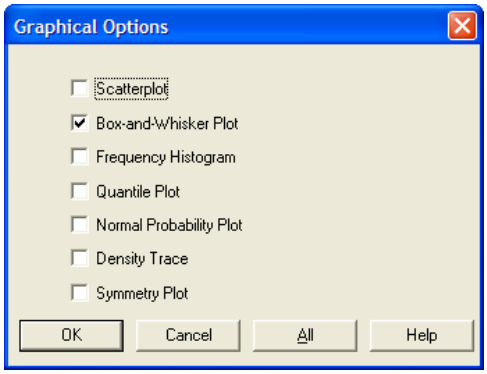
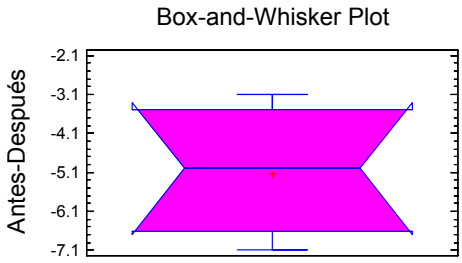


Figura 59. Graphical Options.

RESULTADOS

Paired Samples - Antes & Después

```
Analysis Summary

Data variable: Antes-Después

11 values ranging from -7.1 to -3.1

The StatAdvisor
---------------
   This procedure is designed to test for
significant differences between two data samples
where the data were collected as pairs.  It will
calculate various statistics and graphs for the
differences between the paired data.   Also
included in the procedure are tests designed to
determine whether the mean difference is equal
to zero.
Use the Tabular Options and Graphical Options
buttons on the analysis toolbar to access these
different procedures.
```



Summary Statistics for Antes-Después

Count = 11
Average = -5.12727
Variance = 2.19618
Standard deviation = 1.48195
Minimum = -7.1
Maximum = -3.1
Range = 4.0
Stnd. skewness = 0.165666
Stnd. kurtosis = -1.06793

The StatAdvisor

This table shows summary statistics for Antes-Después. It includes measures of central tendency, measures of variability, and measures of shape. Of particular interest here are the standardized skewness and standardized kurtosis, which can be used to determine whether the sample comes from a normal distribution. Values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate any statistical test regarding the standard deviation. In this case, the standardized skewness value is within the range expected for data from a normal distribution. The standardized kurtosis value is within the range expected for data from a normal distribution.

Confidence Intervals for Antes-Después

95.0% confidence interval for mean:
-5.12727 +/- 0.995591 [-6.12286,-4.13168]
95.0% confidence interval for standard deviation: [1.03547,2.60072]

The StatAdvisor

This pane displays 95.0% confidence intervals for the mean and standard deviation of Antes-Después. The classical interpretation of these intervals is that, in repeated sampling, these intervals Hill contain the true mean or standard deviation of the population from which the data come 95.0% of the time. In practical terms, we can state with 95.0% confidence that the true mean Antes-Después is somewhere between -6.12286 and -4.13168, while the true Standard deviation is somewhere between 1.03547 and 2.60072.

Both intervals assume that the population from which the simple comes can be represented by a normal distribution. While the confidence interval for the mean is quite robust and not very sensitive to violations of this assumption, the confidence interval for the standard deviation is quite sensitive. If the data do not come from a normal distribution, the interval for the Standard deviation may be incorrect. To check whether the data come from a normal distribution, select Summary Statistics from the list of Tabular Options, or choose Normal Probability Plot from the list of Graphical Options.

Hypothesis Tests for Antes-Después

Sample mean = -5.12727
Sample median = -5.0

t-test

Null hypothesis: mean = 0.0
Alternative: not equal

Computed t statistic = -11.4749
P-Value = 4.44472E-7

Reject the null hypothesis for alpha = 0.05.

sign test

Null hypothesis: median = 0.0
Alternative: not equal

Number of values below hypothesized median: 11
Number of values above hypothesized median: 0

Large sample test statistic = 3.01511 (continuity correction applied)
P-Value = 0.00256896

Reject the null hypothesis for alpha = 0.05.

signed rank test

Null hypothesis: median = 0.0
Alternative: not equal

Average rank of values below hypothesized median: 6.0
Average rank of values above hypothesized median: 0.0

Large sample test statistic = 2.8896 (continuity correction applied)
P-Value = 0.00385742

Reject the null hypothesis for alpha = 0.05.

The StatAdvisor

This pane displays the results of three tests concerning the center of the population from which the sample of Antes-Después comes. The first test is a t-test of the null hypothesis that the mean Antes-Después equals 0.0 versus the alternative hypothesis that the mean Antes-Después is not equal to 0.0. Since the P-value for this test is less than 0.05, we can reject the null hypothesis at the 95.0% confidence level. The second test is a sign test of the null hypothesis that the median Antes-Después equals 0.0 versus the alternative hypothesis that the median Antes-Después is not equal to 0.0. It is based on counting the number of

```
values above and below the hypothesized median.
Since the P-value for this test is less than
0.05, we can reject the null hypothesis at the
95.0% confidence level.
The third test is a signed rank test of the null
hypothesis that the median Antes-Después equals
0.0 versus the alternative hypothesis that the
median Antes-Después is not equal to 0.0. It is
based on comparing the average ranks of values
above and below the hypothesized median.  Since
the P-value for this test is less than 0.05, we
can reject the null hypothesis at the 95.0%
confidence level. The sign and signed rank tests
are less sensitive to the presence of outliers
but are somewhat less powerful than the t-test
if the data all come from a single normal
distribution.
```

INTERPRETACIÓN

El punto central es la significación de la diferencia

```
Hypothesis Tests for Antes-Después
Sample mean = -5.12727
Sample median = -5.0
t-test
------
Null hypothesis: mean = 0.0
Alternative: not equal

Computed t statistic = -11.4749
P-Value = 4.44472E-7

Reject the null hypothesis for alpha = 0.05.
```

Al dar clic al botón derecho sobre la ventana de contraste de hipótesis y seleccionar **Pane Options**, se puede seleccionar **Less Than**, para que la hipótesis alterna sea realmente de "menor que"

P-value es muchísimo menor de 0.05 (realmente es 2.22236×10^{-7} = 0.000000222236), entonces se tiene evidencia estadística que el programa de ejercicio si beneficia a los pacientes.

Esto se corrobora con los resultados del intervalo de confianza.

```
Confidence Intervals for Antes-Después
--------------------------------------
95.0% confidence interval for mean:
 -5.12727 +/- 0.995591   [-6.12286,-4.13168]
95.0% confidence interval for standard
deviation: [1.03547,2.60072]
```

El intervalo de confianza para la diferencia promedio no incluye el valor de cero, de aquí que si hay efecto del programa de ejercicio.

Pero (nunca falta un pero …), qué tan válidas son estas conclusiones, ya que esta prueba se basa en la normalidad de los datos.

De los resultados de Summary Statistics se puede retomar la siguiente indicación.

```
Of particular interest here are the standardized
skewness and standardized kurtosis, which can be
used to determine whether the sample comes from
a normal distribution. Values of these
statistics outside the range of -2 to +2
indicate significant departures from normality,
which would tend to invalidate any statistical
test regarding the standard deviation. In this
case, the standardized skewness value is within
the range expected for data from a normal
distribution. The standardized kurtosis value is
within the range expected for data from a normal
distribution.
```

Esto implica que los datos cumplen son el supuesto de normalidad y por lo tanto las conclusiones son totalmente válidas desde el punto de vista estadístico.

NOTA: Recuerdar que se pueden pedir los valores de Sesgo y Curtosis directamente de las Pane Options en la ventana de resultados del Summary.

También se puede tener un análisis visual de esta diferencia o del comportamiento de los valores de antes y después, para lo cual se sigue la siguiente secuencia:

Compare – Multiple Samples – Multiple-Sample Comparison

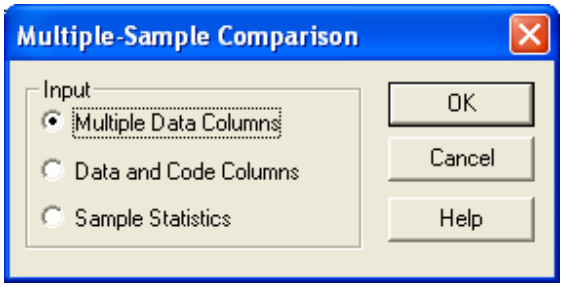


Figura 60. Multiple-Sample Comparison.

Se introducen las variables como se indica a continuación y seleccionar OK.

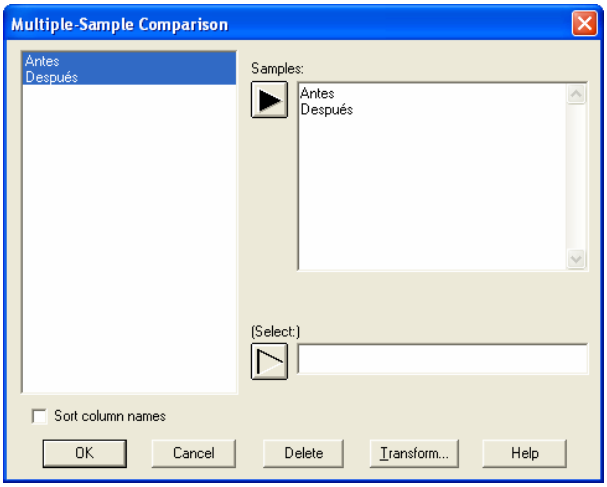
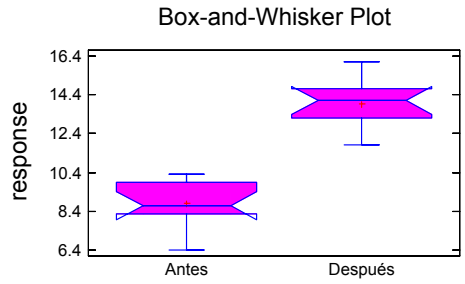


Figura 61. Multiple-Sample Comparison



EJERCICIOS

En cada uno de los siguientes ejercicios establezca claramente el par de hipótesis a contrastar antes de realizarlos

1. Se estudia la vida de anaquel de una bebida carbonatada, para esto se seleccionan 10 botellas al azar y se prueban obteniendo los siguientes resultados en días.

108, 124, 124, 106, 115, 138, 163, 159, 134, 139

Se quiere demostrar que la vida media de anaquel excede los 120 días.

2. El nivel máximo aceptable de exposición a radiación de microondas en Estados Unidos se ha establecido en un promedio de 10 microwatios por centímetro cuadrado. Se teme que un gran transmisor de televisión pueda contaminar el aire del entorno inmediato elevando el nivel de radiación de microondas por encima del límite de seguridad. Si la siguiente es una muestra aleatoria de nueve observaciones (en microwatios por centímetro cuadrado), tomadas en lugares próximos al transmisor.

9, 11, 14, 10, 10, 12, 13, 8, 12

¿Qué conclusión práctica puede extraerse de estos datos?

3. Se utilizan dos máquinas para llenar botellas de plástico con un volumen neto de 16.0 onzas. Puede suponerse que el proceso de llenado es normal, con desviaciones estándar de $\sigma_1 = 0.015$ y $\sigma_2 = 0.018$. El departamento de ingeniería de calidad sospecha que ambas máquinas llenan el mismo volumen neto, sin importar si este volumen es 16.0 onzas o no. Se realiza un experimento tomando muestras aleatorias del llenado de cada máquina. (Diseño y análisis de experimentos, Montgomery C. D., 2002, 2ª. Edición, Limusa Wiley, ejercicio 2-9, pág. 55)

Máquina 1		Máquina 2	
16.03	16.01	16.02	16.03
16.04	15.96	15.97	16.04
16.05	15.98	15.96	16.02
16.05	16.02	16.01	16.01
16.02	15.99	15.99	16.00

Realizar una propuesta de análisis, explicando los criterios para seleccionar cada prueba y explicitando las hipótesis estadísticas a trabajar.

4. Se plantea un estudio para probar si un programa de entrenamiento de relajación disminuye la severidad de los ataques de asma. Para esto se consideran 5 pacientes que sufren este tipo de trastorno respiratorio y se registra la severidad de los síntomas midiendo la dosis de medicamento que requieren para superar un ataque de asma. Tanto al inicio del entrenamiento de una semana como después de él.

Paciente	Inicio	Después del entrenamiento
1	9.0	4.0
2	4.0	1.0
3	5.0	5.0
4	4.0	0.0
5	5.0	1.0

¿Sirve o no sirve el entrenamiento de relajación?

Sugerencia: cuidar el supuesto de normalidad en los datos.

5. Se quiere probar si una nueva actividad mejora la capacidad de lectura en alumnos de primaria. Para esto se aplica la actividad a un grupo de 21 alumnos y se toma como control a otros 23 alumnos a quienes no se les aplica la actividad.

Con actividad		**Sin actividad**	
24	53	42	46
43	56	43	10
58	59	55	17
71	52	26	60
43	62	62	53
49	54	37	42
61	57	33	37
44	33	41	42
67	46	19	55
49	43	54	28
57		20	48
		85	

¿Se tiene evidencia de que la actividad realmente mejora la capacidad de lectura?

Realizar el análisis e interpretar. Reportando en un archivo Word.

CAPÍTULO 5

BONDAD DE AJUSTE Y PRUEBA DE INDEPENDENCIA

El objetivo de la bondad de ajuste es comprobar la hipótesis nula de que un conjunto de datos se extrae o se "ajusta" a una distribución específica de probabilidad.

Esta prueba es muy interesante y útil, ya que permite comparar si un conjunto de datos se ajusta a una normal (o a cualquier otra distribución de interés). Recordando que la normalidad es un supuesto implícito en todas las pruebas inferenciales que se han revisado hasta el momento.

En otras palabras, el juego de hipótesis a trabajar es:

Ho: Los datos se apegan a una distribución normal
Ha: Los datos no se ajustan a una distribución normal.

Ejemplo 1

Se prueba un nuevo fármaco para su posible utilización en el tratamiento de las náuseas asociadas con mareos causados por movimientos. La variable en estudio, Y, es el tiempo en minutos necesarios para obtener alivio. Contrastar Y con la normalidad con base en las siguientes observaciones (Estadística para Biología y Ciencias de la Salud, Milton, J. S. y J. O. Tsokos, Ed. Interamericana MacGraw-Hill, 1987, pág. 407).

2.2	1.5	3.4	4.4	3.9
3.4	4.3	3.7	2.6	4.2
2.5	3.1	3.2	1.9	3.7
3.3	3.8	3.9	3.5	3.1
4.8	4.7	3.4	3.2	3.3
3.0	3.1	4.5	3.8	4.1
3.7	3.5	3.3	2.9	3.0
4.1	3.1	3.6	3.2	2.6

SOLUCIÓN

Ho: Los datos se ajustan a una distribución normal
Ha: Los datos no se ajustan a una distribución normal.

0.- Generar un archivo con una sola columna, de nombre tiempo, y 40 datos.

1. Seguir la secuencia

Describe -> Distributions -> Distribution Fitting (Uncensored Data)

2. Colocar la variable tiempo en el diálogo Data

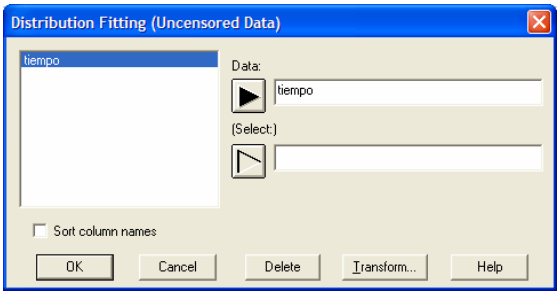


Figura 62. Ajustando a una distribución.

3. Presionar el botón OK y listo.

4. Seleccionar las opciones tabulares, asegurándose de activar **Test for Normality**

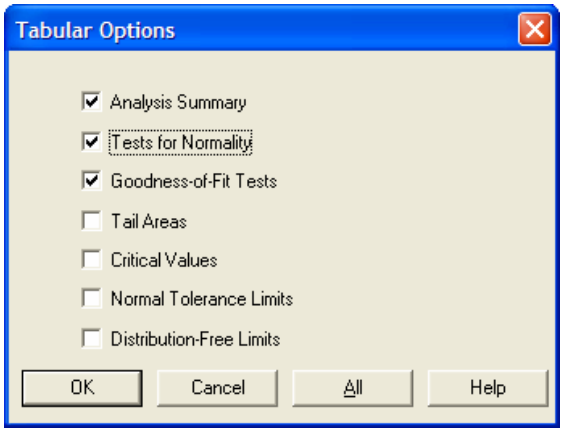


Figura 63. Opciones Tabulares.

5. Seleccionar las opciones gráficas

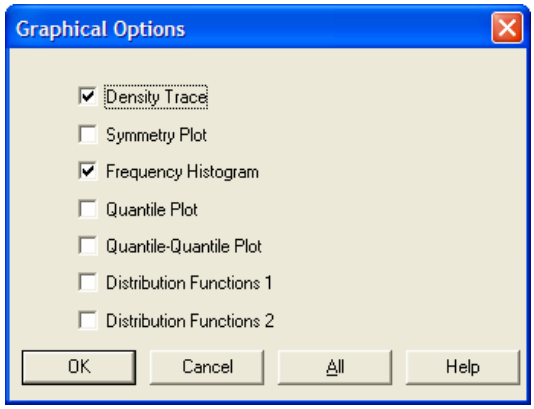


Figura 64. Opciones Gráficas.

RESULTADOS

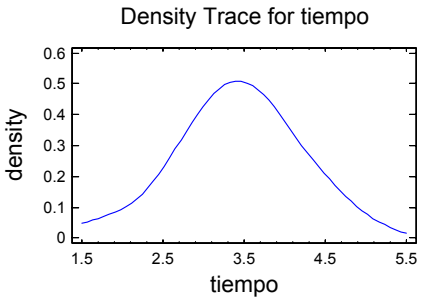
Uncensored Data - tiempo

```
Analysis Summary
Data variable: tiempo
```

40 values ranging from 1.5 to 4.8
Fitted normal distribution:
mean = 3.4125
standard deviation = 0.71439

The StatAdvisor

This analysis shows the results of fitting a normal distribution to the data on tiempo. The estimated parameters of the fitted distribution are shown above. You can test whether the normal distribution fits the data adequately by selecting Goodness-of-Fit Tests from the list of Tabular Options. You can also assess visually how well the normal distribution fits by selecting Frequency Histogram from the list of Graphical Options. Other options within the procedure allow you to compute and display tail areas and critical values for the distribution. To select a different distribution, press the alternate mouse button and select Analysis Options.



Tests for Normality for tiempo

Computed Chi-Square goodness-of-fit statistic = 10.15
P-Value = 0.751137

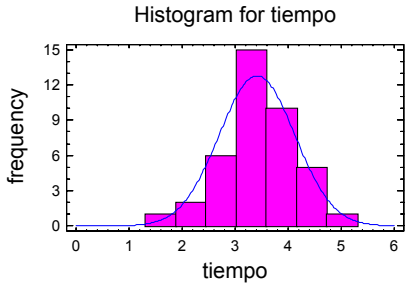
Shapiro-Wilks W statistic = 0.976247
P-Value = 0.666039

Z score for skewness = 0.751035
P-Value = 0.45263

Z score for kurtosis = 0.895745
P-Value = 0.370387

The StatAdvisor

This pane shows the results of several tests run to determine whether tiempo can be adequately modeled by a normal distribution.
The chi-square test divides the range of tiempo into 17 equally probable classes and compares the number of observations in each class to the number expected. The Shapiro-Wilks test is based upon comparing the quantiles of the fitted normal distribution to the quantiles of the data. The standardized skewness test looks for lack of symmetry in the data. The standardized kurtosis test looks for distributional shape which is either flatter or more peaked than the normal distribution.
The lowest P-value amongst the tests performed equals 0.370387.
Because the P-value for this test is greater than or equal to 0.10, we can not reject the idea that tiempo comes from a normal distribution with 90% or higher confidence.



Goodness-of-Fit Tests for tiempo

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	2.64984	6	5.71	0.01
2.64984	3.00819	3	5.71	1.29
3.00819	3.2839	7	5.71	0.29
3.2839	3.5411	8	5.71	0.91
3.5411	3.81681	6	5.71	0.01
3.81681	4.17516	4	5.71	0.51
above 4.17516		6	5.71	0.01

Chi-Square = 3.04975 with 4 d.f.
P-Value = 0.549534

Estimated Kolmogorov statistic DPLUS = 0.056983
Estimated Kolmogorov statistic DMINUS = 0.106828
Estimated overall statistic DN = 0.106828
Approximate P-Value = 0.751313

EDF Statistic	Value	Modified Form	P-Value
Kolmogorov-Smirnov D	0.106828	0.688927	>=0.10*
Anderson-Darling A^2	0.311272	0.317546	0.5380*

*Indicates that the P-Value has been compared to tables of critical values specially constructed for fitting the currently selected distribution. Other P-values are based on general tables and may be very conservative.

The StatAdvisor

This pane shows the results of tests run to determine whether tiempo can be adequately modeled by a normal distribution. The chi-square test divides the range of tiempo into nonoverlapping intervals and compares the number of observations in each class to the number expected based on the fitted distribution. The Kolmogorov-Smirnov test computes the maximum distance between the cumulative distribution of tiempo and the CDF of the fitted normal distribution. In this case, the maximum distance is 0.106828. The other EDF statistics compare the empirical distribution function to the fitted CDF in different ways.
Since the smallest P-value amongst the tests performed is greater than or equal to 0.10, we can not reject the idea that tiempo comes from a normal distribution with 90% or higher confidence.

En cualquier ventana de resultados se puede dar un clic derecho y accesar a Analysis Options, donde se

despliega una lista de 24 posibles distribuciones que se pueden ajustar mediante esta opción.

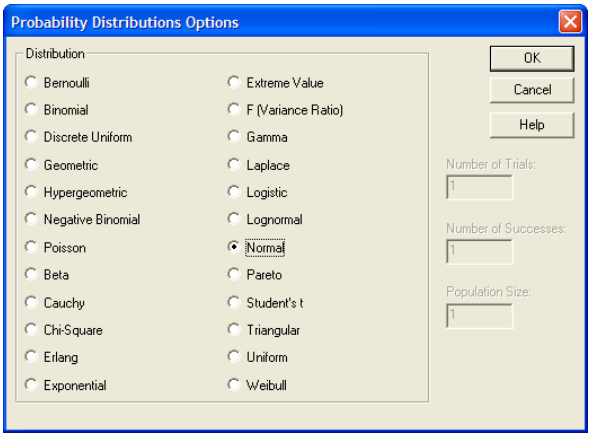


Figura 65. Opciones de Distribuciones de probabilidad.

INTERPRETACIÓN

Sólo hay que recordar cuál es la hipótesis de trabajo, ya que en todas las pruebas se tiene evidencia para no rechazar Ho.

Se tiene evidencia para afirmar que estos datos se apegan a una distribución normal.

PRUEBA DE INDEPENDENCIA

Como ya se vio, la Ji-cuadrada se puede utilizar para comparar una varianza con un valor dado, realizar pruebas de bondad de ajuste (¿hasta dónde una muestra se comporta de acuerdo a una distribución dada?), pero mucha de su popularidad viene de la posibilidad de **probar la independencia** o relación entre dos variables, generalmente de tipo categórico, y arregladas en tablas de doble entrada con **r** filas o renglones y **c** columnas, a la cual se le conoce como tabla de contingencia.

Pasos para hacer una prueba de contingencia:

1. Plantear el juego de hipótesis

Ho: La variable de la columna es independiente de la variable del renglón
Ha: La variable columna NO es independiente de la variable renglón

En términos coloquiales: **Ho: NO hay relación entre las variables.**

2. Tomar una muestra aleatoria y anotar las frecuencias observadas para cada celda de la tabla de contingencias

3. Aplicar la siguiente ecuación: $E_{ij} = \frac{(r_{i\bullet})(c_{\bullet j})}{n}$ (total del renglon i multiplicado por el total de la columna j y dividir este resultado entre el total o tamaño de la muestra), para calcular las frecuencias esperadas en cada celda.

4. Obtener un valor para el estadístico ji-cuadrada

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum \frac{O_{ij}^2}{E_{ij}} - n$$

una medida de la desviación entre las frecuencias observadas y las esperadas.

5. Aplicar la regla de decisión solamente de cola superior.

Manejo numérico.

Tabla de Contingencia, valores observados

	Col 1	Col 2	Col 3	Total
Reng 1	12	2	6	20
Reng 2	6	6	8	20
Total	18	8	14	n = 40

Tabla de Contingencia, valores esperados

	Col 1	Col 2	Col 3	Total
Reng 1	O_{11} = 12 E_{11} = 9	O_{12} = 2 E_{12} = 4	O_{13} = 6 E_{13} = 7	$r_{1\bullet}$ = 20
Reng 2	O_{21} = 6 E_{21} = 9	O_{22} = 6 E_{22} = 4	O_{23} = 8 E_{23} = 7	$r_{2\bullet}$ =20
Total	$c_{\bullet 1}$ =18	$c_{\bullet 2}$ = 8	$c_{\bullet 2}$ = 14	n = 40

Para este caso:

$$\chi^2 = \sum \frac{O_{ij}^2}{E_{ij}} - n = 44.2857 - 40 = 4.2857$$

Con grados de libertad g.l. = (2-1)(3-1) = 2, se tiene una significancia de 0.14, lo que implica NO rechazar Ho.

Una de las primeras preguntas es que tan grande es este valor de ji-cuadrada, por lo que el valor máximo se obtiene con: $\chi^2_{\max} = n(a-1)$, con a es valor más pequeño de las hileras o columnas. En este caso el valor numérico es 40(2-1) = 40(1) = 40.

De aquí se puede obtener el coeficiente de determinación, como: $\frac{\chi^2}{\chi^2_{\max}} = \frac{4.2857}{40} \approx 0.11$. De donde se concluye que el grado de asociación es de únicamente un 11%.

Ejemplo 2

Un grupo de investigadores cree que el número de aciertos, logrados por estudiantes universitarios (FES Zaragoza) en un examen sobre la situación política del país, depende de la carrera que cursa. Para probar su hipótesis aplicaron un cuestionario con 52 preguntas sobre el tema referido a 494 estudiantes, obteniéndose los siguientes datos:

	ACIERTOS				
CARRERA	0-12	13-25	26-38	39-52	Total
Biólogo	43	46	36	25	150
QFB	36	45	39	54	174
I. Q.	40	55	42	33	170
Total	119	146	117	112	494

A partir de estos datos, ¿podría decirse con un nivel de significación del 5% que el grupo de investigación está en lo cierto? Enuncie y contraste la hipótesis apropiada

SOLUCIÓN

0. Ho: El número de aciertos no depende de la carrera.
   Ha: El número de aciertos depende de la carrera.

1. Crear un archivo con cinco columnas, una para identificar las filas, llamada carrera (de tipo **character**), y cuatro numéricas (A, B. C y D).

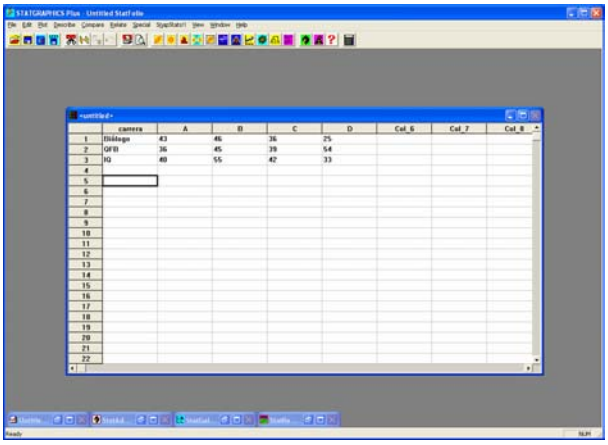


Figura 66. Tablas de contingencia o de clasificación cruzada.

2. Siga la secuencia

Describe → Categorical Data → Contingency Table

(note que los datos van en **COLUMNS** y el identificador de filas en **LABELS**.

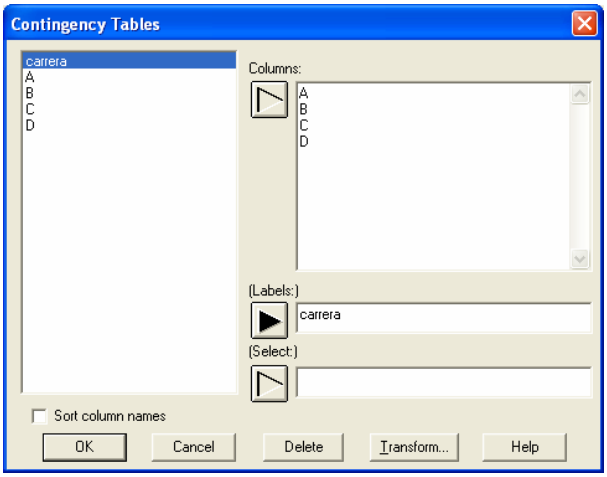


Figura 67. Tablas de contingencia.

3. Seleccione las opciones tabulares y escoja

Analysis Summary, Frecuency Table y Chi-Square

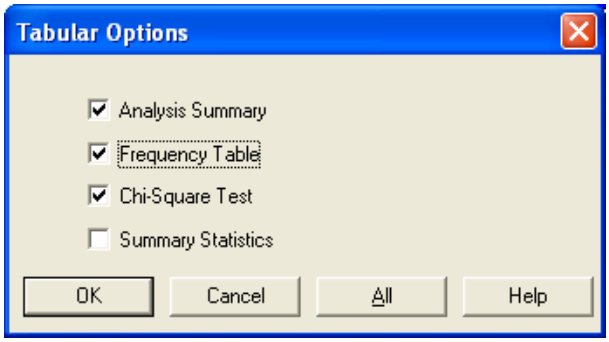


Figura 68. Opciones tabulares.

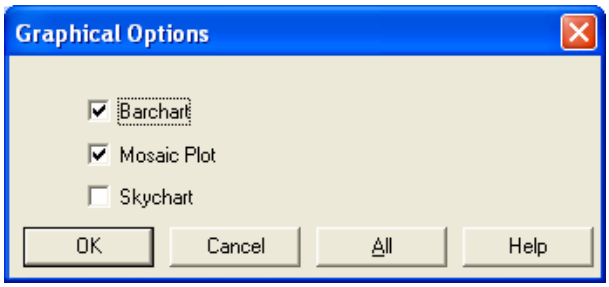


Figura 69. Opciones gráficas.

RESULTADOS

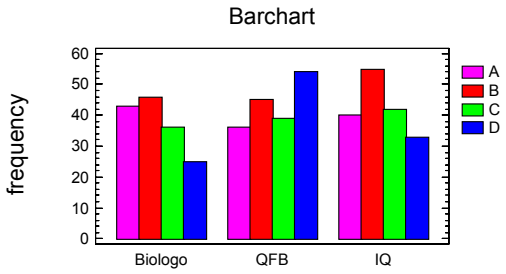
```
Analysis Summary

Column variables:
     A
     B
     C
     D

Number of observations: 494
Number of rows: 3
Number of columns: 4


The StatAdvisor
---------------
   This procedure constructs various statistics
and graphs for a two-way table.  Of particular
interest is the test for independence between
rows and columns, which you can run by choosing
Chi-Square Test on the list of Tabular Options.
```

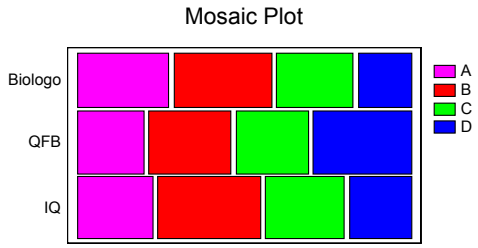
```
Frequency Table

                                          Row
          A       B       C       D     Total
-----------------------------------------------
Biologo | 43    | 46    | 36    | 25    |   150
        | 8.70% | 9.31% | 7.29% | 5.06% | 30.36%
-----------------------------------------------
QFB     |    36 |    45 |    39 |    54 |   174
        | 7.29% | 9.11% | 7.89% | 10.93%| 35.22%
-----------------------------------------------
IQ      |    40 |    55 |    42 |    33 |   170
        | 8.10% | 11.13%| 8.50% | 6.68% | 34.41%
-----------------------------------------------
Column      119     146     117     112      494
Total    24.09%  29.55% 23.68%  22.67%  100.00%

Cell contents:
    Observed frequency
    Percentage of table
```

```
The StatAdvisor
---------------
   This table displays counts for a 3 by 4
table.  The first number in each cell of the
table is the count or frequency.  The second
number shows the percentage of the entire table
represented by that cell. For example, there
were 43 values in the first row and first
column. This represents 8.70445% of the 494
values in the table.
```

```
Chi-Square Test
-----------------------------------------
     Chi-Square        Df         P-Value
-----------------------------------------
          12.23         6          0.0571
-----------------------------------------
```

```
The StatAdvisor
---------------
   The chi-square test performs a hypothesis
test to determine whether or not to reject the
idea that the row and column classifications are
independent. Since the P-value is less than
0.10, we can reject the hypothesis that rows and
columns are independent at the 90% confidence
level. Therefore, the observed row for a
particular case is related
to its column.
```





INTERPRETACIÓN:

La interpretación se realiza con base en el p-value = 0.0571 < 0.10. Se debe resaltar que se rechaza Ho, esto implica que si hay relación del número de aciertos con la carrera.

Otro aspecto a notar es que, por omisión se trabaja con α=0.10.

Ejemplo 3

Se realiza un estudio para determinar si hay alguna asociación aparente entre el peso de un muchacho y un éxito precoz en la escuela, a juicio de un Psicólogo escolar. Se toma una muestra aleatoria de 500 estudiantes y se clasifican con base en dos criterios, el peso y el éxito en la escuela

	Sobrepeso	**Sin sobrepeso**
Con éxito	162	263
Sin éxito	38	37

Ho: Tener sobrepeso es independiente del éxito en la escuela.

Ha: Tener sobrepeso no es independiente del éxito en la escuela.

SOLUCIÓN

0.- Crear un archivo con 3 columnas, una para identificar las filas, llamada éxito, de tipo carácter y dos numéricas, una para sobrepeso y otra para sin sobrepeso. De tal manera que el archivo tiene el siguiente aspecto.

exito	sobrepeso	no sobre
Con	162	263
Sin	38	37

1. Seguir la secuencia

Describe -> Categorical Data -> Contingencia Table

2. Colocar las variables en el sitio adecuado del diálogo que se despliega.

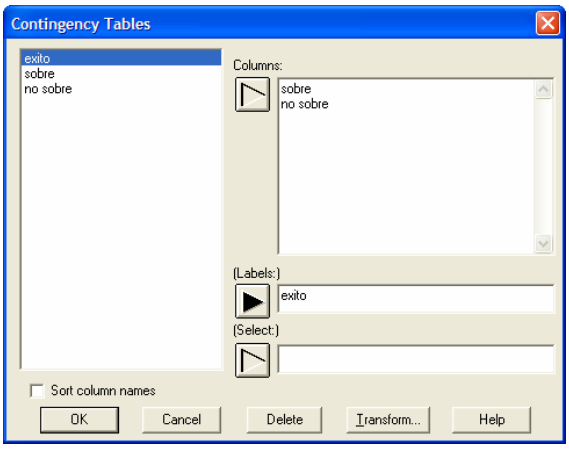


Figura 70. Tablas de contingencia.

Observe que los datos van en **Columns** y el identificador de filas o renglones en **Labels.**

3. Dar un clic en OK para ver los resultados parciales.

4. Seleccionar las opciones tabulares

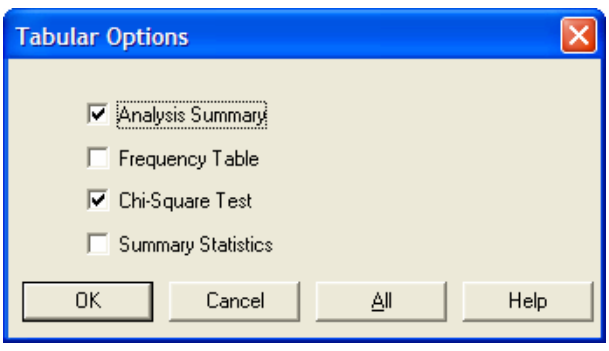


Figura 71. Opciones Tabulares.

5. Seleccionar las opciones gráficas

En este caso son más importantes los resultados numéricos, pero si requiere de algún apoyo visual se deben explorar las diferentes posibilidades.

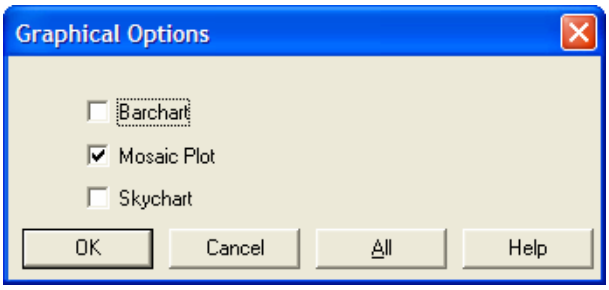


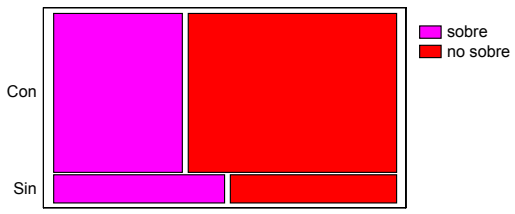
Figura 72. Opciones Gráficas.

RESULTADOS

Contingency Tables

```
Analysis Summary

Column variables:
     sobre
     no sobre

Number of observations: 500
Number of rows: 2
Number of columns: 2


The StatAdvisor
---------------
   This procedure constructs various statistics
and graphs for a two-way table.  Of particular
interest is the test for independence between
rows and columns, which you can run by choosing
Chi-Square Test on the list of Tabular Options.
```

Mosaic Plot




```
Frequency Table

                                          Row
             sobre        no sobre        Total
           ---------------------------
Con        |       162 |       263 |       425
           |    32.40% |    52.60% |    85.00%
           ---------------------------
Sin        |        38 |        37 |        75
           |     7.60% |     7.40% |    15.00%
           ---------------------------
Column             200          300          500
Total           40.00%       60.00%      100.00%
```

```
The StatAdvisor
---------------
   This table displays counts for a 2 by 2
table.  The first number in each cell of the
table is the count or frequency.  The second
number shows the percentage of the entire table
represented by that cell.
For example, there were 162 values in the first
row and first column. This represents 32.4% of
the 500 values in the table.
```

```
Chi-Square Test
-----------------------------------
Chi-Square  Df    P-Value
-----------------------------------
    4.18    1    0.0408
    3.68    1    0.0552 (with Yates' correction)
-----------------------------------------
```

```
The StatAdvisor
---------------
   The chi-square test performs a hypothesis
test to determine whether or not to reject the
idea that the row and column classifications are
independent.  Since the P-value is less than
0.10, we can reject the hypothesis that rows and
columns are independent at the 90% confidence
level.  Therefore, the observed row for a
particular case is related to its column.

NOTE: the P-value with Yates' correction was
used because it should be more accurate for a 2-
by-2 table.
```

```
Summary Statistics
-------------------------------------------------
                          With Rows  With Columns
Statistic     Symmetric   Dependent    Dependent
-------------------------------------------------
Lambda           0.0036      0.0000       0.0050
Uncertainty Coeff.  0.0075      0.0097       0.0061
Somer's D       -0.0871     -0.0667      -0.1255
Eta                         0.0915       0.0915
-------------------------------------------------
Statistic          Value       P-Value       Df
-------------------------------------------------
Contingency Coeff.   0.0911
Cramer's V           0.0915
Conditional Gamma   -0.2502
Pearson's R         -0.0915      0.0205      498
Kendall's Tau b     -0.0915      0.0410
Kendall's Tau c     -0.0640
-------------------------------------------------
```

```
The StatAdvisor
---------------
   The statistics shown here measure the degree
of association between rows and columns.  Of
particular interest are the contingency
coefficient and lambda, which measure the degree
of association on a scale of 0 to 1.  Lambda
measures how useful the row (or column) factor
is in predicting the other factor.  For example,
the value of lambda with columns dependent
equals 0.005.  This means that there is a 0.5%
reduction in error when rows are used to predict
columns.  For those statistics with P values, P
values less than 0.05 indicate a significant
association between rows and columns at the 95%
confidence level.
```

INTERPRETACIÓN

```
Since the P-value is less than 0.10, we can
reject the hypothesis that rows and columns are
independent at the 90% confidence level.
Therefore, the observed row for a particular
case is related to its column.
```

La interpretación se realiza con base en el p-value = 0.0408 < 0.10. Se debe resaltar que se rechaza Ho, esto implica que si hay relación del sobrepeso con el éxito escolar.
Sin embargo, como se trata de una tabla 2×2 aplica una corrección de Yates cuyo p-value es 0.0552 < 0.10, llegaándose a la misma conclusión.

Otro aspecto a notar es: por omisión se trabaja con α =0.10.

```
- Of particular interest are the contingency
coefficient and lambda, which measure the degree
of association on a scale of 0 to 1. Lambda
measures how useful the row (or column) factor
is in predicting the other factor.
```

Esta última opción permite trabajar también con medidas de asociación (con interpretación semejante a un coeficiente de correlación) entre filas y columnas.

Para el ejemplo se deben "teclear" 500 datos en dos columnas una para sobrepeso y otra para éxito, cada una de ellas con dos posibles resultados, por ejemplo 1 para tiene sobrepeso y 0 para no tiene sobrepeso. De manera semejante 1 para éxito y 0 para no éxito.

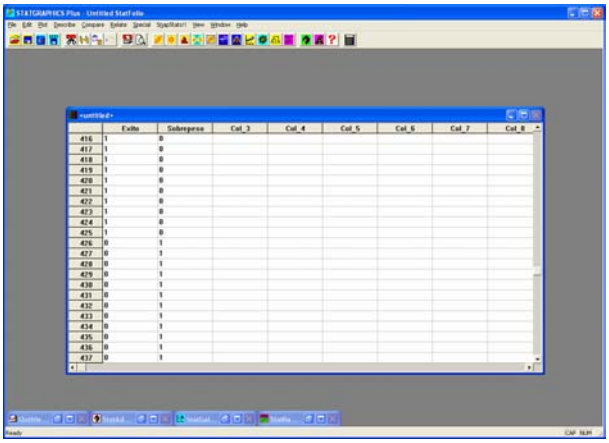


Figura 73. Datos sin condensar.

NOTA; NOTA, NOTA: Para datos sin condensar se trabaja la secuencia.

Describe -> Categorical Data -> Crosstabulation

En el diálogo que aparece definir qué variable va en las columnas y cuál en las filas o renglones.

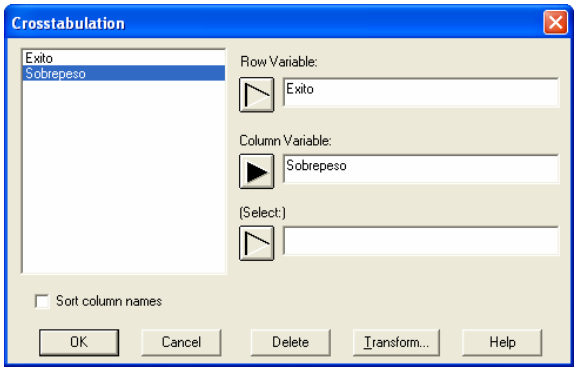


Figura 74. Clasificación cruzada.

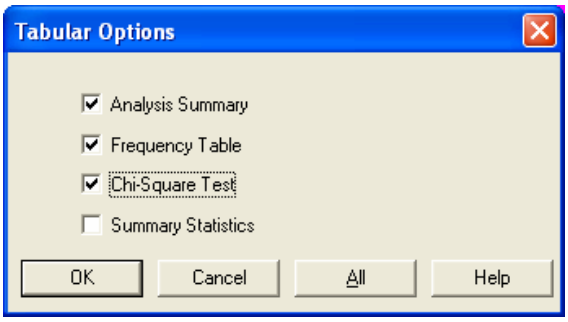


Figura 75. Opciones Tabulares.

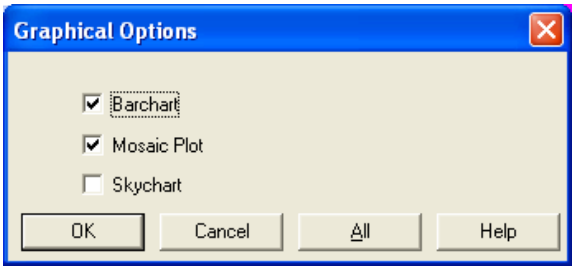
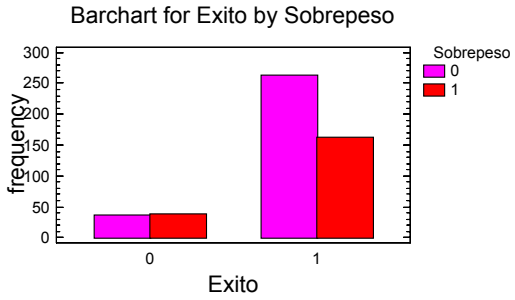


Figura 76. Opciones Gráficas.

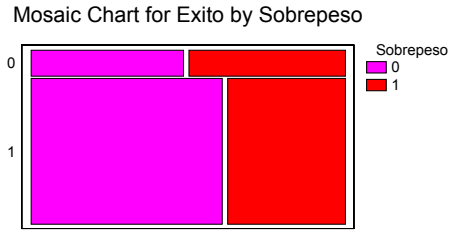
RESULTADOS

Crosstabulation - Exito by Sobrepeso

```
Analysis Summary

Row variable: Exito
Column variable: Sobrepeso

Number of observations: 500
Number of rows: 2
Number of columns: 2

The StatAdvisor
---------------
   This procedure constructs a two-way table
showing the frequency of occurrence of unique
pairs of values for Exito and Sobrepeso. It
constructs a 2 by 2 contingency table for the
data and displays the results in various ways.
Of particular interest is the test for
independence between rows and columns, which you
can run by choosing Chi-Square Test on the list
of Tabular Options.
```




```
Frequency Table for Exito by Sobrepeso
                                        Row
        0             1             Total
-----------------------------------------
0   |      37   |      38   |         75
    |   7.40%   |   7.60%   |     15.00%
-----------------------------------------
1   |     263   |     162   |        425
    |  52.60%   |  32.40%   |     85.00%
-----------------------------------------
Column   300          200          500
Total 60.00%      40.00%      100.00%

Cell contents:
    Observed frequency
    Percentage of table

The StatAdvisor
---------------
   This table shows how often the 2 values of
Exito occur together with each of the 2 values
of Sobrepeso.  The first number in each cell of
the table is the count or frequency.  The second
number shows the percentage of the entire table
represented by that cell.  For example, there
were 37 times when Exito equaled 0 and Sobrepeso
equaled 0.  This represents 7.4% of the total of
500 observations.
```




```
Chi-Square Test
-----------------------------------------
Chi-Square  Df   P-Value
-----------------------------------------
  4.18       1    0.0408
  3.68       1    0.0552 (with Yates'
correction)
-----------------------------------------
```

```
The StatAdvisor
---------------
   The chi-square test performs a hypothesis
test to determine whether or not to reject the
idea that the row and column classifications are
independent. Since the P-value is less than
0.10, we can reject the hypothesis that rows and
columns are independent at the 90% confidence
level. NOTE: the P-value with Yates' correction
was used because it should be more accurate for
a 2-by-2 table. Therefore, the observed value of
Exito for a particular case is related to its
value for Sobrepeso.
```

INTERPRETACIÓN

Al igual que en el caso anterior, al ser una tabla 2 × 2 se aplica la corrección de Yates cuyo p-value es 0.0552 < 0.10, llegándose a la misma conclusión.

Ejemplo 4

La siguiente tabla muestra los resultados de una encuesta realizada a 15 estudiantes, a quienes se les interrogó acerca de si tenían o no miedo de caminar alrededor del campo universitario por la noche.

	Género	
Miedo	Hombre	Mujer
Sí	6	2
No	1	6

Determine si hay relación entre el género y el tener miedo de caminar por el campo universitario de noche. Utilice la prueba la Exacta de Fisher.

SOLUCIÓN

Ho: El miedo es independiente del género.
Ha: El miedo no es independiente del género.

0.- Crear un archivo con 3 columnas, una para identificar las filas, llamada miedo, de tipo carácter y dos numéricas, una para hombre y otra para mujer. De tal manera que el archivo tiene el siguiente aspecto.

Miedo	hombre	mujer
Sí	6	2
No	1	6

1. Seguir la secuencia

Describe -> Categorical Data -> Contingencia Table

2. Colocar las variables en el sitio adecuado del diálogo que se despliega.

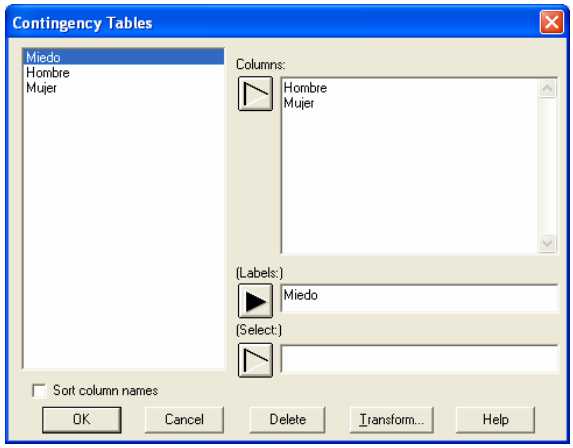


Figura 77. Tablas de contingencia.

Observe que los datos van en **Columns** y el identificador de filas o renglones en **Labels.**

3. Dar un clic en OK para ver los resultados parciales.

4. Seleccionar las opciones tabulares

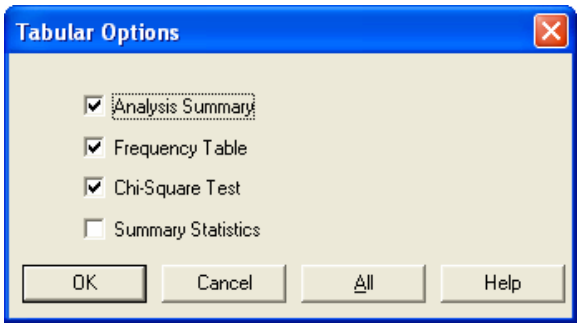
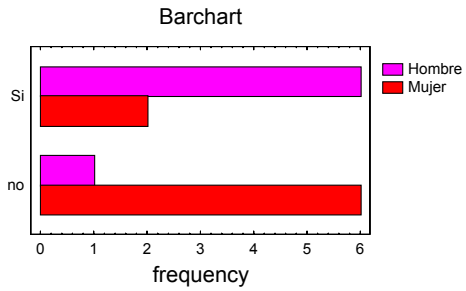


Figura 78. Tablas de contingencia.

RESULTADOS

Contingency Tables

```
Analysis Summary

Column variables:
     Hombre
     Mujer

Number of observations: 15
Number of rows: 2
Number of columns: 2

The StatAdvisor
---------------
   This procedure constructs various statistics
and graphs for a two-way table.  Of particular
interest is the test for independence between
rows and columns, which you can run by choosing
Chi-Square Test on the list of Tabular Options.
```



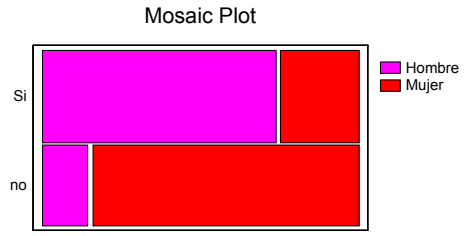
Frequency Table

	Hombre	Mujer	Row Total
Si	6	2	8
	40.00%	13.33%	53.33%
no	1	6	7
	6.67%	40.00%	46.67%
Column Total	7 46.67%	8 53.33%	15 100.00%

Cell contents:
Observed frequency
Percentage of table

The StatAdvisor

This table displays counts for a 2 by 2 table. The first number in each cell of the table is the count or frequency. The second number shows the percentage of the entire table represented by that cell. For example, there were 6 values in the first row and first column. This represents 40.0% of the 15 values in the table.



Chi-Square Test

Chi-Square	Df	P-Value
5.53	1	0.0187
3.36	1	0.0668 (with Yates' correction)

Warning: some cell counts < 5.

Fisher's Exact Test for 2 by 2 Tables

One-tailed P-value = 0.0317016
Two-tailed P-value = 0.0405594

The StatAdvisor

The chi-square test performs a hypothesis test to determine whether or not to reject the idea that the row and column classifications are independent. Since the P-value is less than 0.10, we can reject the hypothesis that rows and columns are independent at the 90% confidence level. Therefore, the observed row for a particular case is related to its column. NOTE: the P-value with Yates' correction was used because it should be more accurate for a 2-by-2 table. Fisher's exact test was also performed. As with the chi-square test, P-values less than 0.05 indicate a significant relationship between the row and column classifications.

INTERPRETACIÓN

La interpretación se realiza con base en el p-value = 0.0668 < 0.10. Se debe resaltar que se rechaza Ho, esto implica que si hay relación del miedo con el género.

Sin embargo, como se trata de una tabla 2 × 2 con frecuencias muy pequeñas, se aplica la prueba Exacta de Fisher cuyo p-value es 0.040559 < 0.10, llegando a la misma conclusión.

CAPÍTULO 6

EJERCICIOS GENERALES, 1ª PARTE

Antes de continuar es importante hacer un alto y revisar, mediante el desarrollo de algunos ejercicios los avances que se tienen hasta el momento.

INSTRUCCIONES

a. En cada uno de los ejercicios escriba con sus propias palabras el par de hipótesis a probar, para después escribirlas en términos de hipótesis nulas y alternativas.

b. Resolver en la computadora, cuidando el cumplimiento de supuestos.

c. Hacer un reporte en WORD, para su revisión.

1. Las siguientes son medidas de profundidad, en una estación de investigación oceanográfica, medida en metros: 46.8, 43.8, 44.6, 38.9, 45.6, 52.1, 40.1, 53.4, 49.4, 53.2, 46.3, 47.8, 42.2 y 44.9.
   ¿Contradicen estos datos la aseveración de que la profundidad promedio en esta zona es de 42.5 m?

2. Los siguientes son Km por galón obtenidos en 40 tanques llenos de cierta gasolina. De la cual se afirma que su rendimiento promedio es de 34.2 Km por galón.

33.3	34.2	35.2	34.4	34.7	34.2	34.0	34.5
35.2	35.8	34.6	33.6	34.3	34.1	34.5	34.2
34.6	34.9	35.3	35.1	34.6	33.8	34.2	35.3
34.5	34.1	33.8	34.3	35.6	34.7	34.0	334.8
34.1	34.4	33.2	35.0	34.8	35.2	34.5	34.1

   ¿Qué se puede decir del rendimiento promedio?

3. Para determinar la efectividad de un nuevo sistema de control de tránsito, se observó el número de accidentes en diez cruceros peligrosos cuatro semanas antes (medición 1) y cuatro semanas después (medición 2) de la instauración del nuevo sistema, obteniendo los siguientes resultados.

M1	3	4	2	5	3	2	3	6	1	1
M2	1	2	3	2	3	0	2	3	2	0

   ¿Qué se puede decir del nuevo sistema?

4. Las siguientes son calificaciones de 15 estudiantes, en un mismo examen aplicado a la mitad y final de un curso de estadística.

Mitad	66	88	75	90	63	58	75	82
Fin	73	91	78	86	69	67	75	80

Mitad		73	84	85	93	70	82	90
Fin		76	89	81	96	76	90	97

   Los alumnos afirman que mejoraron sus calificaciones, ¿si o no?

5. Los siguientes son números de empleados ausentes, de dos departamentos de una empresa importante, durante 28 días.

Dpto1	4	2	6	3	1	2	5	1	3	6
Dpto2	3	5	6	6	4	4	2	4	4	5

2	7	4	1	2	0	6	4	1	4
5	1	6	3	5	3	5	6	2	1

2	1	4	2	0	5	2	3
4	2	1	4	1	3	3	4

   ¿Hay evidencia que soporte la afirmación del jefe del Departamento 1, de que hay menos ausentismo en su Departamento?

6. Para probar la hipótesis de que el suburbio A es más "caro" que el suburbio B, se tomaron datos del gasto semanal en alimentos, de 10 familias con dos hijos, seleccionadas al azar.

SA	278.60	270.78	270.50	267.89	275.38
SB	262.63	263.12	275.16	275.91	255.35

SA	272.00	286.45	264.19	267.95	271.15
SB	266.51	278.19	263.76	271.72	260.78

7. Se le pregunta a muestras aleatorias independientes de 80 individuos solteros, 120 casados y 100 viudos o divorciados si "los amigos y la vida social", "el trabajo o la actividad preponderante" o "la salud y la condición física" contribuyen más a su felicidad general. Obteniendo los siguientes resultados.

	Solteros	Casados	Viudos o divorciados
Amigos y vida social	41	49	42
Trabajo o actividad preponderante	27	50	33
Salud y condición física	12	21	25
Total	80	120	100

¿Qué se puede decir de estos resultados?

8. Se desea investigar si hay alguna relación entre las calificaciones de una prueba de clasificación de las personas que han participado en un cierto programa de capacitación laboral y su posterior desempeño en el trabajo. Para lo cual se dispone de una muestra aleatoria de 400 casos tomada de registros muy grandes.

	Desempeño		
Calificación	Deficiente	Regular	Bueno
Abajo del promedio	67	64	25
Promedio	42	76	56
Arriba del promedio	10	23	37

9. Los siguientes datos presentan la producción de fríjol (en toneladas por hectárea) en terrenos esencialmente similares, pero probando 4 distancias entre plantas.

D1	D2	D3	D4
23.1	21.7	21.9	19.8
22.8	23.0	21.3	20.4
23.2	22.4	21.6	19.3
23.4	21.1	20.2	18.5
23.6	21.9	21.6	19.1
21.7	23.4	23.8	21.9

Aportan estos datos evidencia de que las variaciones en producciones deben exclusivamente al azar o se deben a la distancia entre plantas.

CAPÍTULO 7

MÉTODOS NO-PARAMÉTRICOS

En la Estadística paramétrica nuestro interés es hacer estimaciones y pruebas de hipótesis acerca de uno o más parámetros de la población o poblaciones. Cuando se utiliza estadística paramétrica se debe tener la precaución de verificar que la población o poblaciones de donde provienen las muestras están distribuidas normalmente, aunque sea en forma aproximada.

Los Métodos no-paramétricos o métodos de distribución libre, en contraste, no dependen del conocimiento de cómo se distribuye la población. De esto se deduce que estos métodos son convenientes si no se conoce la distribución de la población. Otra ventaja es que, por lo general, los cálculos necesarios son más sencillos. Sin embargo, no se puede esperar que en el caso de una cierta distribución, la cantidad de información dada por un método no-paramétrico sea la misma que daría un método paramétrico que sólo se aplica a esa distribución específica. Es decir, si se conoce que la distribución de los datos es normal, una prueba paramétrica es más eficiente que una no-paramétrica.

Los métodos no-paramétricos pueden ser usados para analizar datos de tipo cualitativo, ya sean ordinales (jerarquizados) o nominales; así como también para datos cuantitativos, mientras que los métodos paramétricos sólo se pueden usar para datos cuantitativos (continuos). Las pruebas no-paramétricas pueden utilizarse en problemas cuyos datos provienen de una escala ordinal y, por lo tanto, la única alternativa es tomar a la ***mediana*** como la medida descriptiva más adecuada Por otra parte, los métodos no-paramétricos se utilizan para todo tipo de distribución, no solamente para las normales, por ello utilizan una medida de tendencia central más robusta (menos sensible a valores extremos) como lo es la ***mediana.***

Una desventaja, sin embargo, de los métodos no-paramétricos es que no son aplicables a experimentos complejos en los cuales se manejan muchas variables, mientras que los métodos paramétricos como el Análisis de Varianza es comúnmente utilizado en estas situaciones.

Al igual que en la estadística paramétrica, en la estadística no-paramétrica existen métodos para realizar inferencias con una, dos o más muestras.
Las pruebas para una muestra son fundamentalmente la prueba del signo de la mediana y la del Rango con Signo de Wilcoxon.

Las pruebas con dos muestras van a depender de si las muestras son pareadas o independientes, para el caso de muestras pareadas se utiliza la prueba del Rango con Signo de Wilcoxon y para el caso de muestras independientes la prueba de Mann-Whitney.

Con más de dos muestras también, va a depender si las muestras son independientes en cuyo caso se utiliza la Prueba de Kruskal-Wallis y si son dependientes o en bloques, la prueba de Friedman.

Cabe aclarar que el STATGRAPHICS es un paquete diseñado para realizar pruebas paramétricas, aunque incluye algunos de los métodos noparamétricos, no los incluye todos. Para estos métodos es mejor el SPSS.

¿Cómo realizar estas pruebas con STATGRAPHICS? Esto lo veremos con ejemplos.

Ejemplo 1

En los ovarios de la rata se pueden obtener folículos en diferentes etapas de desarrollo, los cuales se clasifican en folículos de reserva, medianos, grandes y pre-ovulatorios. Un inves-tigador desea saber si la medición del número de folículos de reserva en los ovarios de la rata es de 45, como lo reporta alguna literatura. Para ello el investigador sacrificó 10 ratas al azar y contó en un ovario de rata el número de folículos de reserva. Con un α = 0.05 pruebe la hipótesis del investigador usando la prueba del Signo de la Mediana y la del Rango con Signo de Wilcoxon.

Rata	# de folículos
1	53
2	65
3	45
4	44
5	29
6	48
7	61
8	62
9	32
10	40

SOLUCIÓN:

0) Hipótesis: *Ho: Md* = 45 *Ha: Md* ≠45

1) Generar un archivo con los datos

2) Seguir la secuencia:

Describe → Numeric Data → One-variable Analysis

3) Colocar la variable folículos en el diálogo **DATA**

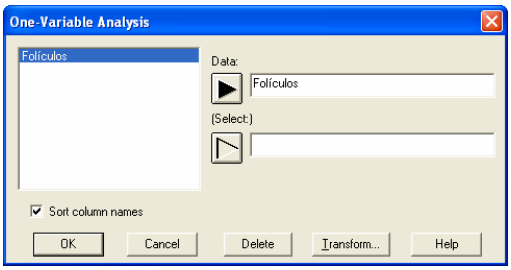


Figura 79. Diálogo para introducir la variable.

4) Presionar el botón OK

5) Seleccione las opciones tabulares, asegúrese de seleccionar: **Hypothesis Tests**

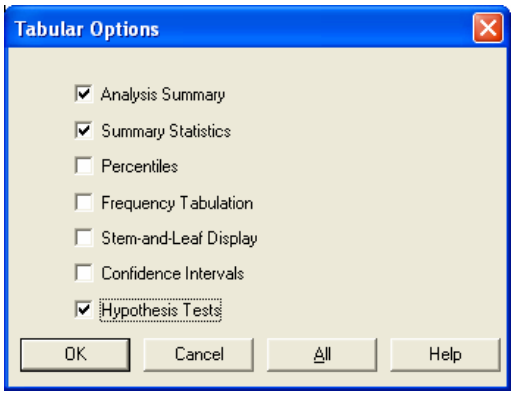


Figura 80. Opciones Tabulares.

6) Sobre la pantalla de Hipótesis Tests, dé clic al botón derecho y seleccione
Pane Options

7) Cambie el valor predeterminado de la media de 0 a 45

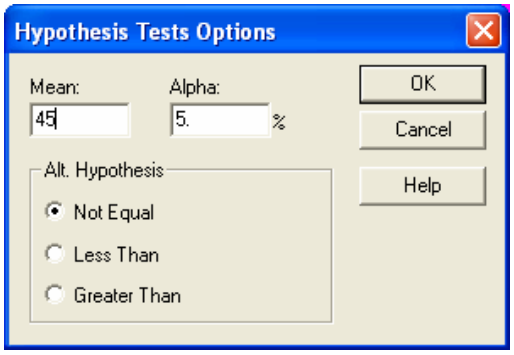


Figura 81. Pane Options.

RESULTADOS

NOTA: el paquete muestra muchas pruebas entre ellas las paramétricas y las noparamétricas, sin embargo, las que nos interesan son sólo las noparamétricas, por ello, éstas se presentan sombreadas

One-Variable Analysis - Folículos

```
Analysis Summary

Data variable: Folículos

10 values ranging from 29.0 to 65.0

The StatAdvisor
---------------
   This procedure is designed to summarize a
single sample of data. It will calculate various
statistics and graphs. Also included in the
procedure are confidence intervals and
hypothesis tests. Use the Tabular Options and
Graphical Options buttons on the analysis
toolbar to access these different procedures.

Summary Statistics for Folículos

Count = 10
Average = 47.9
Median = 46.5
Mode =
Variance = 153.878
Standard deviation = 12.4047
Minimum = 29.0
Maximum = 65.0
Range = 36.0
Stnd. skewness = -0.109295
Stnd. kurtosis = -0.713412

The StatAdvisor
---------------
   This table shows summary statistics for
Folículos.  It includes measures of central
tendency, measures of variability, and measures
of shape.  Of particular interest here are the
standardized skewness and standardized kurtosis,
which can be used to determine whether the
sample comes from a normal distribution. Values
of these statistics
outside the range of -2 to +2 indicate
significant departures from normality, which
would tend to invalidate any statistical test
```

```
regarding the standard deviation. In this case,
the standardized skewness value is within the
range expected for data from a normal
distribution. The standardized kurtosis value
is within the range expected for data from a
normal distribution.

Hypothesis Tests for Folículos

Sample mean = 47.9
Sample median = 46.5

t-test
------
Null hypothesis: mean = 45.0
Alternative: not equal

Computed t statistic = 0.739282
P-Value = 0.478573

Do not reject the null hypothesis for alpha =
0.05.

sign test
---------
Null hypothesis: median = 45.0
Alternative: not equal

Number of values below hypothesized median: 4
Number of values above hypothesized median: 5

Large sample test statistic = 0.0 (continuity
correction applied)
P-Value = 0.999994

Do not reject the null hypothesis for alpha =
0.05.

signed rank test
----------------
Null hypothesis: median = 45.0
Alternative: not equal

Average rank of values below hypothesized
median: 4.875
Average rank of values above hypothesized
median: 6.9

Large sample test statistic = 0.662972
(continuity correction applied)
P-Value = 0.507346

Do not reject the null hypothesis for alpha =
0.05.
The StatAdvisor
---------------
   This pane displays the results of three tests
concerning the center of the population from
which the sample of Folículos comes. The first
test is a t-test of the null hypothesis that the
mean Folículos equals 45.0 versus the
alternative hypothesis that the mean Folículos
is not equal to 45.0. Since the P-value for this
test is greater than or equal to 0.05, we cannot
reject the null hypothesis at the 95.0%
confidence level. The second test is a sign test
of the null hypothesis that the median Folículos
equals 45.0 versus the alternative hypothesis
that the median Folículos is not equal to 45.0.
It is based on counting the number of values
above and below the hypothesized median. Since
the P-value for this test is greater than or
equal to 0.05, we cannot reject the null
hypothesis at the 95.0% confidence level. The
third test is a signed rank test of the null
hypothesis that the median Folículos equals 45.0
versus the alternative hypothesis that the
median Folículos is not equal to 45.0. It is
based on comparing the average ranks of values
above and below the hypothesized median. Since
the P-value for this test is greater than or
equal to 0.05, we cannot reject the null
hypothesis at the 95.0% confidence level. The
sign and signed rank tests are less sensitive to
the presence of outliers but are somewhat less
powerful than the t-test if the data all come
from a single normal distribution.
```

INTERPRETACIÓN

La mediana no es significativamente diferente de 45 ya que p = 0.999994 > 0.05, usando la prueba del signo de la mediana. Igualmente p-value = 0.507346 > 0.05 para la prueba del rango con signo de Wlcoxon.

Ejemplo 2

En 15 pares de parcelas se plantan dos variedades de soya y los rendimientos observados se registran en el siguiente cuadro. Verifique que esta información conduce al no rechazo de la hipótesis de que los rendimientos son iguales usando la prueba del Rango con Signo de Wilcoxon. Use 5% de nivel de significación.

Par	Var. Soya I	Var Soya II
1	135	134
2	129	137
3	130	151
4	146	142
5	127	138
6	128	142
7	125	140
8	151	122
9	151	121
10	128	138
11	134	122
12	132	119
13	121	130
14	136	139
15	121	128

SOLUCIÓN:

0) Hipótesis: *Ho:* $Md_1 - Md_2 = 0$ *Ha:* $Md_1 - Md_2 \neq 0$

1) Generar un archivo con dos columnas de datos.

2) Seguir la secuencia:

   Compare → Two Samples → Pair-Sample Comparison

3) Colocar las variables en los diálogos

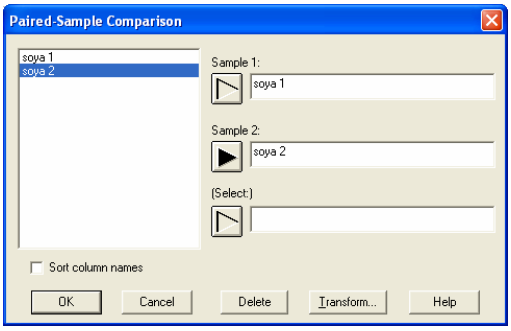


Figura 82. Comparación de muestras pareadas.

4) Presionar el botón OK

5) Seleccione las opciones tabulares, asegúrese de seleccionar: **Hypothesis Tests**

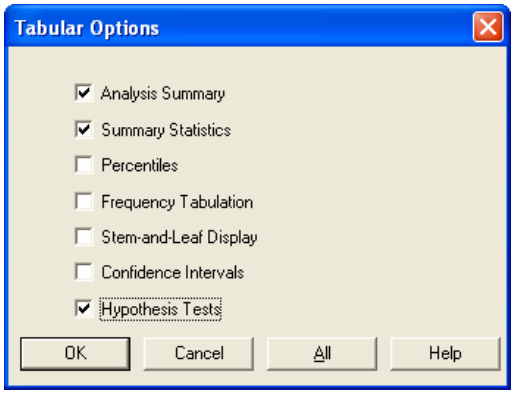


Figura 83. Opciones Tabulares.

RESULTADOS

Paired Samples - soya 1 & soya 2

```
Analysis Summary

Data variable: soya 1-soya 2

15 values ranging from -21.0 to 30.0

The StatAdvisor
---------------
   This procedure is designed to test for
significant differences between two data samples
where the data were collected as pairs. It will
calculate various statistics and graphs for the
differences between the paired data.  Also
included in the procedure are tests
designed to determine whether the mean
difference is equal to zero. Use the Tabular
Options and Graphical Options buttons on the
analysis toolbar to access these different
procedures.

Summary Statistics for soya 1-soya 2
Count = 15
Average = -0.6
Variance = 239.4
Standard deviation = 15.4726
Minimum = -21.0
Maximum = 30.0
Range = 51.0
Stnd. skewness = 1.48086
Stnd. kurtosis = 0.0244126


The StatAdvisor
---------------
   This table shows summary statistics for soya
1-soya 2.  It includes measures of central
tendency, measures of variability, and measures
of shape.  Of particular interest here are the
standardized skewness and standardized kurtosis,
which can be used to determine whether the
sample comes from a normal distribution. Values
of these statistics
outside the range of -2 to +2 indicate
significant departures from normality, which
would tend to invalidate any statistical test
regarding the standard deviation. In this case,
the standardized skewness value is within the
range expected for data from a normal
distribution.  The standardized kurtosis value
is within the range expected for data from a
normal distribution.

Hypothesis Tests for soya 1-soya 2

Sample mean = -0.6
Sample median = -7.0

t-test
------
Null hypothesis: mean = 0.0
Alternative: not equal

Computed t statistic = -0.150188
P-Value = 0.882759

Do not reject the null hypothesis for alpha =
0.05.

sign test
---------
Null hypothesis: median = 0.0
Alternative: not equal

Number of values below hypothesized median: 9
Number of values above hypothesized median: 6

Large sample test statistic = 0.516398
(continuity correction applied)
P-Value = 0.605574

Do not reject the null hypothesis for alpha =
0.05.

signed rank test
----------------
Null hypothesis: median = 0.0
Alternative: not equal

Average rank of values below hypothesized
median: 7.55556
Average rank of values above hypothesized
median: 8.66667
```

```
Large sample test statistic = 0.425971
(continuity correction applied)
P-Value = 0.670125

Do not reject the null hypothesis for alpha =
0.05.

The StatAdvisor
---------------
   This pane displays the results of three tests
concerning the center of the population from
which the sample of soya 1-soya 2 comes. The
first test is a t-test of the null hypothesis
that the mean soya 1-soya 2 equals 0.0 versus
the alternative hypothesis that the mean soya 1-
soya 2 is not equal to 0.0.  Since the P-value
for this test is greater than or equal to 0.05,
we cannot reject the null hypothesis at the
95.0% confidence level. The second test is a
sign test of the null hypothesis that the median
soya 1-soya 2 equals 0.0 versus the alternative
hypothesis that the median soya 1-soya 2 is not
equal to
0.0. It is based on counting the number of
values above and below the hypothesized median.
Since the P-value for this test is greater than
or equal to 0.05, we cannot reject the null
hypothesis at the 95.0% confidence level.  The
third test is a signed rank test of the null
hypothesis that the median soya 1-soya 2 equals
0.0 versus the alternative hypothesis that the
median soya 1-soya 2 is not equal to 0.0. It is
based on comparing the average ranks of values
above and below the hypothesized median.  Since
the P-value for this test is greater than or
equal to 0.05, we cannot reject the null
hypothesis at the 95.0% confidence level. The
sign and signed rank tests are less sensitive to
the presence of outliers but are somewhat less
powerful
than the t-test if the data all come from a
single normal distribution.
```

INTERPRETACIÓN:

Con la prueba del signo de la mediana, se tiene un p-value de 0.605574 > 0.10 y con la prueba del rango con signo de Wilcoxon un p-value de 0.670125 >0.10. En ambos casos no se rechaza Ho y se concluye que las medianas de las variedades no son significativamente diferentes.

Ejemplo 3

Las siguientes son medidas de albúmina (*g*/100*mL*) de 17 personas normales y 13 personas hospitali-zadas. ¿Concluiría usted que las dos poblaciones son diferentes al 5% de nivel de significación? Use la prueba del signo de la mediana y la prueba de Mann-Whitney y compare los resultados.

Albúmina (*g/100mL*)			
Normales		***Hospitalizadas***	
2.4	3.0	1.5	3.1
3.5	3.2	2.0	1.3
3.1	3.5	3.4	1.5
4.0	3.8	1.7	1.8
4.2	3.9	2.0	2.0
3.4	4.0	3.8	1.5
4.5	3.5	3.5	
5.0	3.6		
2.9			

SOLUCIÓN:

0) *Ho:* $Md_N - Md_H = 0$
Ha: $Md_N - Md_H \neq 0$

1) Generar un archivo con dos columnas y 30 renglones, una que identifique el tipo de paciente y otra con los datos.

2) Seguir la secuencia:

Compare → Two Samples → Two-Sample Comparison

3) Colocar la variable **albúmina** en el diálogo data e individuos en simple code. Seleccionar Imput: Data and Code Columns

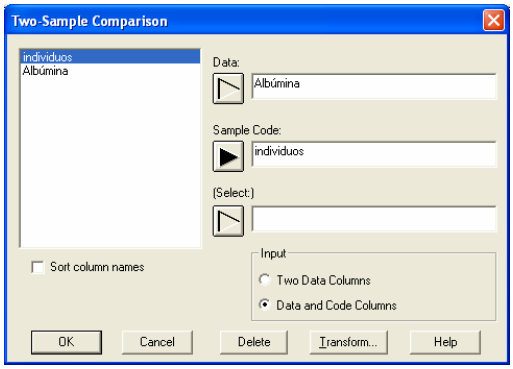


Figura 84. Comparación de dos muestras independientes.

NOTA: Los datos se pueden ingresar en dos columnas, una para “antes” y otra para “después”, en cuyo caso se debe activar la opción **Two Data_Columns** en **Input**.

4) Presionar el botón OK

5) Seleccione las opciones tabulares, asegúrese de seleccionar: **Comparison of medians**

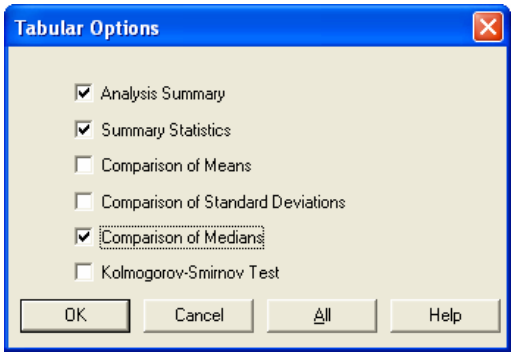


Figura 85. Opciones Tabulares.

RESULTADOS

Two-Sample Comparison - Albúmina & individuos

```
Analysis Summary for Albúmina

Sample 1: individuos=1
Sample 2: individuos=2

Sample 1: 17 values ranging from 2.4 to 5.0
Sample 2: 13 values ranging from 1.3 to 3.8

The StatAdvisor
---------------
   This procedure is designed to compare two
samples of data. It will calculate various
statistics and graphs for each sample, and it
will run several tests to determine whether
there are statistically significant differences
between the two samples.

Summary Statistics for Albúmina

        individuos=1        individuos=2
----------------------------------------
Count         17                  13
Median         3.5                 2.0
Mode           3.5
Minimum        2.4                 1.3
Maximum        5.0                 3.8
Range          2.6                 2.5
----------------------------------------

The StatAdvisor
---------------
   This table shows summary statistics for the
two samples of data. Other tabular options
within this analysis can be used to test whether
differences between the statistics from the two
samples are statistically significant. Of
particular interest here are the standardized
skewness and standardized kurtosis, which can be
used to determine whether the samples come from
normal distributions. Values of these statistics
outside the range of -2 to +2 indicate
significant departures from normality, which
would tend to invalidate the tests which compare
the standard deviations. In this case, both
standardized skewness values are within the
range expected. Both standardized kurtosis
values are within the range expected.
Comparison of Medians for Albúmina
---------------------------------

Median of sample 1: 3.5
Median of sample 2: 2.0

Mann-Whitney (Wilcoxon) W test to compare
medians

   Null hypothesis: median1 = median2
   Alt. hypothesis: median1 NE median2

   Average rank of sample 1: 20.4118
   Average rank of sample 2: 9.07692

   W = 27.0   P-value = 0.000497351

The StatAdvisor
---------------
   This option runs a Mann-Whitney W test to
compare the medians of the two samples. This
test is constructed by combining the two
samples, sorting the data from smallest to
largest, and comparing the average ranks of the
two samples in the combined data. Since the
P-value is less than 0.05, there is a
statistically significant difference between the
medians at the 95.0% confidence level.
```

INTERPRETACIÓN

Las medianas de los dos grupos son significativamente diferentes ya que p = 0.00049 < 0.10.

Ejemplo 4

Se midieron los tiempos de cristalización de 3 tipos diferentes de amalgama, en segundos, en 12 molares

Velvalley	*Katalley*	*Dispersalley*
92	102	89
99	99	91
98	93	87
91	106	
97		

¿Se puede concluir que los 3 tipos de amalgamas difieren respecto al tiempo de cristalización?

SOLUCIÓN:

0) *Ho:* Los tres tipos de amalgamas tienen el mismo tiempo de cristalización.
   Ha: Los tres tipos de amalgamas tienen diferentes tiempos de cristalización

1) Generar un archivo con dos columnas y 12 renglones, una que identifique el tipo de amalgama y otra con los datos.

2) Seguir la secuencia:

Compare → Multiple Samples → Multiple-Sample Comparison

3) Seleccionar Imput: Data and Code Columns

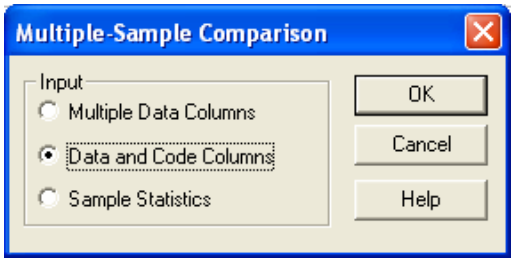


Figura 86. Opción de ingreso de los datos en las comparaciones de múltiples muestras (Kruskal-Wallis).

4) Presionar el botón OK

5) Coloque la variable **tiempo** en el diálogo data y amalgama en simple code.

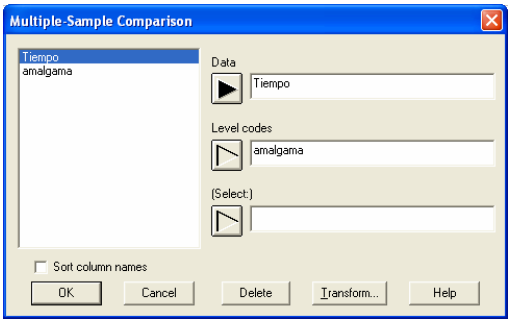


Figura 87. Comparaciones de múltiples muestras.

6) Presionar el botón OK

7) Seleccione las opciones tabulares, asegúrese de seleccionar: **Kruskal-Wallis Test**

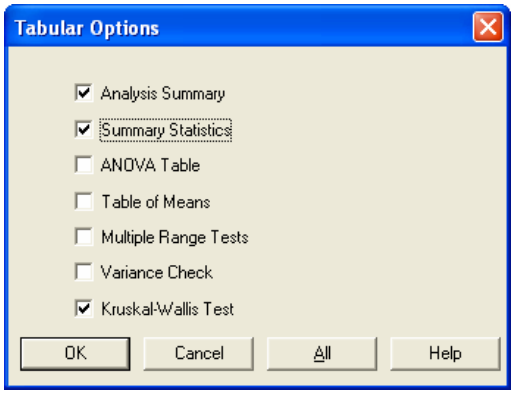


Figura 88. Opciones tabulares.

8) Selecciones las opciones gráficas, marque la opción de **Box and Whisker Plot**

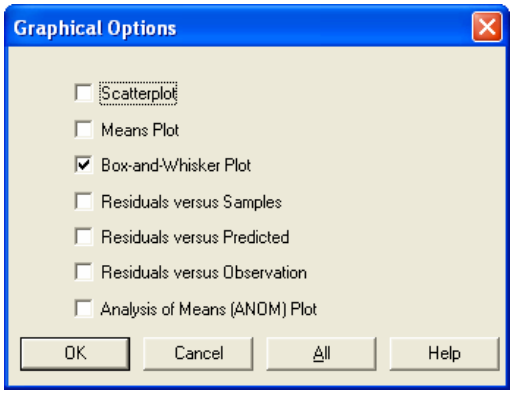


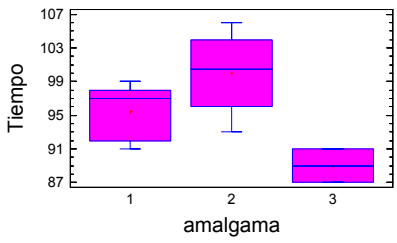
Figura 89. Opciones gráficas.

RESULTADOS

Multiple-Sample Comparison

```
Analysis Summary
Dependent variable: Tiempo
Factor: amalgama
Number of observations: 12
Number of levels: 3

The StatAdvisor
---------------
   This procedure compares the data in 3 columns
of the current data file.  It constructs various
statistical tests and graphs to compare the
samples.  The F-test in the ANOVA table will
test whether there are any significant
differences amongst the means.  If there are,
the Multiple Range Tests will tell you which
means are significantly different from which
others.  If you are worried about the presence
of outliers, choose the Kruskal-Wallis Test
which compares medians instead of means.  The
various plots will help you judge the practical
significance of the results, as well as allow
you to look for possible violations of the
assumptions underlying the analysis of variance.
```

Box-and-Whisker Plot




```
Summary Statistics for Tiempo
amalgama  Count  Average Median Variance
----------------------------------------
1             5     95.4   97.0   13.3
2             4    100.0  100.5   30.0
3             3     89.0   89.0    4.0
----------------------------------------
Total        12     95.3333 95.0   32.6061
```

```
amalgama Std dev   Min   Max   Range Std. skewness
---------------------------------------
1        3.64692  91.0  99.0    8.0
2        5.47723  93.0 106.0   13.0
3        2.0      87.0  91.0    4.0
---------------------------------------
Total    5.71017  87.0 106.0   19.0

amalgama    Stnd skewness   Stnd. kurtosis
---------------------------------------
1           -0.4404          -1.30126
2           -0.397523         0.10433
3            0.0
---------------------------------------
Total        0.486309        -0.48026

The StatAdvisor
---------------
   This table shows various statistics for each
of the 3 columns of data.    To test for
significant differences amongst the column
means, select Analysis of Variance from the list
of Tabular Options.  Select Means Plot from the
list of Graphical Options to display the means
graphically.
```

```
Kruskal-Wallis Test for Tiempo by amalgama

amalgama  Sample Size  Average Rank
---------------------------------------
1             5             6.6
2             4             9.625
3             3             2.16667
---------------------------------------
Test statistic = 7.39369
P-Value = 0.0248016
```

```
The StatAdvisor
---------------
   The Kruskal-Wallis test tests the null
hypothesis that the medians within each of the 3
columns is the same.  The data from all the
columns is first combined and ranked from
smallest to largest.  The average rank is then
computed for the data in each column.  Since the
P-value is less than 0.05, there is a
statistically significant difference amongst the
medians at the 95.0% confidence level.    To
determine which medians are significantly
different from which others, select Box-and-
Whisker Plot from the list of Graphical Options
and select the median notch option.
```

INTERPRETACIÓN

Las tres amalgamas tienen tiempos de cristalización diferentes porque p= 0.0248 < 0.05

Ejemplo 5

Se determinó la presión osmótica de una solución a distintas concentraciones por tres métodos *A, B* y *C*. Los datos obtenidos se muestran en la siguiente tabla. Se desea saber si los tres métodos son iguales o diferentes.

Concen-	*Método*		
	A	***B***	***C***
1	2.59	2.40	2.36
2	5.06	4.81	4.63
3	7.61	7.21	6.80
4	10.14	9.62	8.90
5	12.75	12.00	10.90
6	15.39	14.40	12.80
7	18.13	16.80	14.70
8	20.91	19.20	16.50
9	23.72	21.60	18.20
10	26.64	24.00	19.80

SOLUCIÓN:

0) Hipótesis:
Ho: Los tres métodos son iguales
Ha: Los tres métodos son diferentes

1) Generar un archivo con tres columnas y 10 renglones, una para la presión correspondiente a cada método. Observe que en este ejemplo se quieren comparar los tres métodos, medidos en cada concentración, por lo que se trata de un diseño de bloques al azar (Friedman)

2) Seguir la secuencia:

Compare → Multiple Samples → Multiple-Sample Comparison

3) Seleccionar Input: Múltiple Data Columns

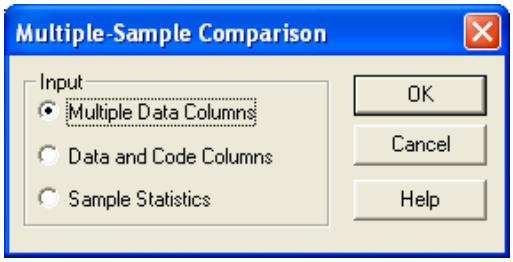


Figura 90. Comparaciones de múltiples muestras

4) Presionar OK

5) Seleccionat las tres columnas e introducirlas en el diálogo samples

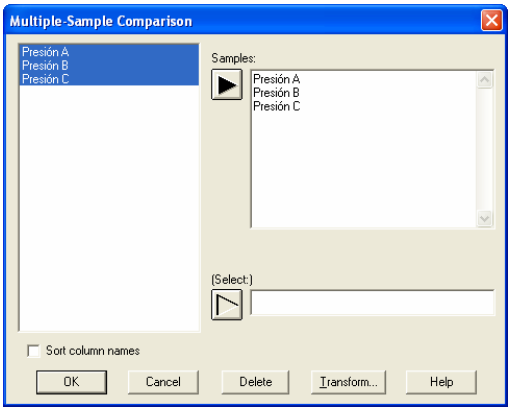


Figura 91. Comparaciones de múltiples muestras.

6) Presionar OK

7) Ir a Tabular Options y seleccionar Kruskal-Wallis and Friedman Tests

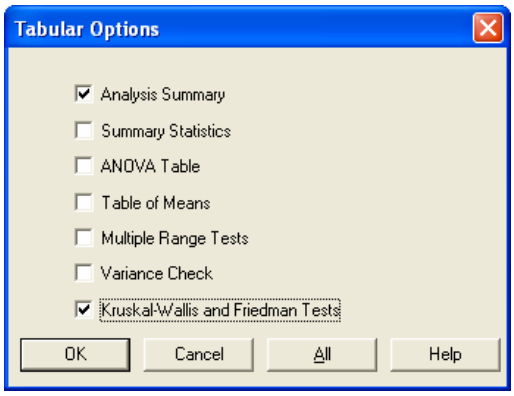


Figura 92. Opciones tabulares.

8) Presionar OK

9) Dar clic con el botón derecho y seleccionar Pane Options y en Rank Tests Options, seleccionar Friedman Test

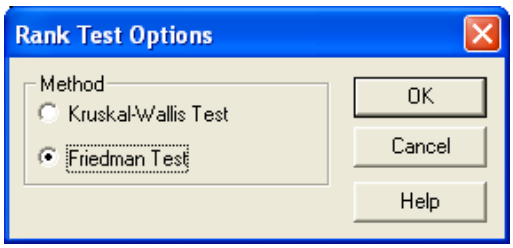


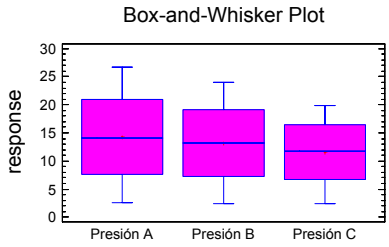
Figura 93. Selección de la prueba de Friedman.

10) Presionar OK

RESULTADOS

Multiple-Sample Comparison

```
Analysis Summary

Sample 1: Presión A
Sample 2: Presión B
Sample 3: Presión C

Sample 1: 10 values ranging from 2.59 to 26.64
Sample 2: 10 values ranging from 2.4 to 24.0
Sample 3: 10 values ranging from 2.36 to 19.8


The StatAdvisor
---------------
   This procedure compares the data in 3 columns
of the current data file.  It constructs various
statistical tests and graphs to compare the
samples.  The F-test in the ANOVA table will
test whether there are any significant
differences amongst the means.  If there are,
the Multiple Range Tests will tell you which
means are significantly different from which
others.  If you are worried about the presence
of outliers, choose the Kruskal-Wallis Test
which compares medians instead of means.  The
various plots will help you judge the practical
significance of the results, as well as allow
you to look for possible violations of the
assumptions underlying the analysis of variance.
```




```
Friedman Test

          Sample Size    Average Rank
----------------------------------------
Presión A     10             3.0
Presión B     10             2.0
Presión C     10             1.0
----------------------------------------
Test statistic = 20.0
P-Value = 0.0000453999
```

```
The StatAdvisor
---------------
   The Friedman test tests the null hypothesis
that the medians within each of the 3 columns is
the same.  The data in each row ranked from
smallest to largest. The average rank is then
computed for each column.  Since the P-value is
less than 0.05, there is a statistically
significant difference amongst the medians at
the 95.0% confidence level. To determine which
medians are significantly different from
which others, select Box-and-Whisker Plot from
the list of Graphical Options and select the
median notch option.
```

INTERPRETACIÓN

Los tres métodos para la determinación de la presión son diferentes porque p = 0.000045 < 0.05.

EJERCICIOS

1. Diez estudiantes obtuvieron las siguientes calificaciones en un examen de estadística: 72, 95, 76, 80, 90, 82, 90, 60, 50, 86. Pruebe la hipótesis de que la mediana de las calificaciones es 75. Use la prueba del signo de la mediana y la del rango con signo de Wilcoxon.

2. Para comparar la velocidad de dos marcas de calculadoras en la realización de cálculos estadísticos, un operador experimentado realizó seis operaciones en cada una de las calculadoras de modelos equivalentes. La siguiente tabla muestra el número de segundos que tardó en realizar los mencionados cálculos.

Cálculo	Calc. *A*	*Calc.* B
1	25	23
2	62	75
3	46	56
4	123	167
5	89	95
6	365	429

¿Existe diferencia en la velocidad entre las marcas de las calculadoras? Use α = 0.05 y las pruebas del Signo de la Mediana y la del Rango con Signo de Wilcoxon.

3. Una compañía aseguradora de automóviles comparó los pagos (en dólares) realizados a los propietarios de 8 automóviles y los clasificó según la marca del carro.

Marca *A*	Marca *B*
353	453
597	527
634	568
696	228
813	725
649	523
593	568
658	155

¿Se justifica tener precios diferentes para las diferentes marcas?

4. Cuatro grupos de pacientes de terapia física fueron sometidos a diferentes tratamientos. Al final de un período determinado a cada grupo se le aplicó una prueba para medir la efectividad del tratamiento. Se obtuvieron los siguientes resultados.

I	*II*	*III*	*IV*
64	76	58	95
88	70	74	90
72	90	66	80
80	80	60	87
79	75	82	88
71	82	75	85

¿Presentan los datos suficiente evidencia que indique una diferencia entre los tratamientos?

5. Dieciséis personas obesas participa-ron en un estudio para comparar 4 dietas reductoras. Se agrupó a los sujetos de acuerdo con su peso inicial y cada uno de los 4 sujetos de cada uno de los grupos se asignaron al azar a las dietas. Al final del experimento se pesaron los sujetos. La tabla siguiente muestra la reducción de peso en libras.

Pesos iniciales (libras)	Dietas			
	A	*B*	*C*	*D*
150-174	12	26	24	23
175-199	15	29	23	25
200-225	15	27	25	24
> 225	18	38	33	31

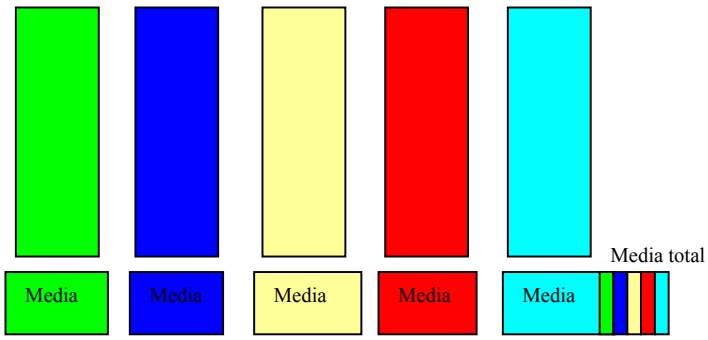
Después de eliminar diferencias debidas a peso inicial, ¿proporcionan estos datos suficiente evidencia que indique una diferencia entre las dietas? Use α = 0.05.

CAPÍTULO 8

ANÁLISIS DE VARIANZA Y DISEÑO DE EXPERIMENTOS

MOTIVACIÓN AL ANÁLISIS DE VARIANZA (ANOVA, ANDEVA o ANVA)

Suponga un experimento donde se quieren comparar 5 tratamientos, para ver si su respuesta promedio es la misma para los 5 o si hay algunas diferentes.



De antemano el investigador asume que hay diferencia, si no que sentido tiene el experimento. También se sabe que en cada tratamiento debe haber un efecto de variaciones debida a la causa que se está controlando (temperatura, presión, etcétera) y una variación debida al azar, la cual es inevitable.

La variación entre tratamientos se mide como una varianza de la media de cada tratamiento con respecto a la gran media.

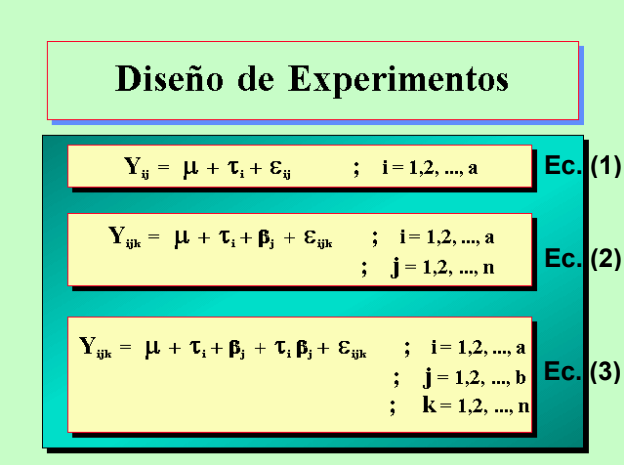
La variación dentro de tratamientos se mide comparando cada observación o medición con respecto a la media del respectivo tratamiento y en términos del análisis de varianza se le conoce como cuadrado medio del error.

Ahora, si se tienen dos varianzas (entre tratamientos y dentro de tratamientos) lo que se puede hacer es compararlas mediante una prueba de F.

$$F = \frac{\text{Varianza entre tratamientos}}{\text{Varianza dentro tratamientos}}$$

La variación dentro de tratamientos se debe al azar y si no se puede establecer diferencia estadística entre estas varianzas, entonces no hay efecto de tratamiento y la variación se debe al azar.

MODELOS MÁS COMUNES EN EL DISEÑO DE EXPERIMENTOS



Diseño Completamente al Azar (DCA), de un factor o One-Way

La característica esencial es que todas las posibles fuentes de variación o de influencia están controladas y sólo hay efecto del factor en estudio. Este es el experimento ideal, todo controlado y lo único que influye es el factor de estudio.

Diseño de Bloques al Azar Completo (DBAC)

Sigue siendo un diseño de una vía pero hay alguna fuente con un gradiente de variación, que influye o afecta en el experimento, por lo tanto hay que cuantificar su efecto y eliminarlo de la varianza dentro de tratamientos, para evitar que nos conduzca a valores bajos de F y se llegue a conclusiones erróneas.

Diseños Factoriales

La tercera ecuación, de la figura anterior, muestra un diseño con dos factores de estudio, donde el mayor interés está en el efecto de la interacción, $\tau_i\beta_j$. Nótese la semejanza entre el modelo de la ecuación 2 y la 3, en la figura anterior.

ANÁLISIS DE VARIANZA DE UN FACTOR O DE UNA VÍA O DISEÑO COMPLETAMENTE AL AZAR.

Para mostrar los cálculos numéricos se tiene el siguiente ejemplo.

Un biólogo decide estudiar los efectos del etanol en el tiempo de sueño. Se seleccionó una muestra de 20 ratas, de edad semejante, a cada rata se le administró una inyección oral con una concentración en particular de etanol por peso corporal. El movimiento ocular rápido (REM) en el tiempo de sueño para cada rata se registró entonces durante un periodo de 24 horas, con los siguientes resultados.

(Modificado de Jay L. Devore, Probabilidad y estadística para ingeniería y ciencias, 5ª. Edición, Ed. Thomson Learning, México, 2001, pág. 412)

	Tratamientos				
	0(control)	1 g/Kg	2 g/Kg	4 g/Kg	
	88.6	63.0	44.9	31.0	
	73.2	53.9	59.5	39.6	
	91.4	69.2	40.2	45.3	
	68.0	50.1	56.3	25.2	
	75.2	71.5	38.7	22.7	
$Y_{\bullet j}$	396.4	307.7	239.6	163.8	$Y_{\bullet\bullet} = 1107.5$
$\overline{Y}_{\bullet j}$	79.28	61.54	47.92	32.76	$\overline{Y}_{\bullet\bullet} = 55.375$

Como primer paso se debe tener en cuenta el par de hipótesis a trabajar.

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Ha: $\mu_i \neq \mu_j$, para algún $i \neq j$

Las fórmulas de cálculo son.

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(Y_{ij} - \overline{Y}..\right)^2 = n\sum_{i=1}^{n}\left(\overline{Y}_{\bullet j} - \overline{Y}..\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{k}\left(Y_{ij} - \overline{Y}_{\bullet j}\right)^2$$

$$SC_{Total} = SC_{trat} + SC_{error}$$

$$SC_{Total} = (88.6-55.375)^2 + (73.2-55.375)^2 + \cdots + (22.7-55.375)^2 = \\ = 73697575$$

$$SC_{Trat} = 5\left[(79.28-55.375)^2 + \cdots + (32.76-55.375)^2\right] = \\ = 5882.357$$

$$SC_{Error} = (88.6-79.28)^2 + (73.2-79.28)^2 + \cdots + (22.7-32.76)^2 = \\ = 1487.4$$

Tratamientos $k = 4$, con repeticiones $n = 5$, por lo tanto $k * n = N = 4 * 5 = 20$

Siguiendo otra estrategia de cálculo:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} Y_{ij}^2 = 88.6^2 + 73.2^2 + \cdots + 22.7^2 = \\ = 68697.57$$

$\frac{Y_{\bullet\bullet}^2}{N} = \frac{1107.5^2}{20} = 61327.8$, valor que se conoce como factor de corrección

$$SC_{Total} = \sum_{i=1}^{n}\sum_{j=1}^{k} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N} = 68697.57 - 61327.8 = \\ = 7369.77$$

$$SC_{Trat} = \frac{\sum_{j=1}^{k} Y_{\bullet j}^2}{n_j} - \frac{Y_{\bullet\bullet}^2}{N} = \frac{396.4^2 + \cdots + 163.8^2}{5} - 61327.8 = \\ = \frac{336050.85}{5} - 61327.8 = 5882.4$$

$$SC_{Error} = SC_{Total} - SC_{Trat} = 7369.77 - 5882.4 = 1487.4$$

Tabla de ANOVA

Fuente de Variación	g.l.	Suma de Cuadrados	Cuadrados Medios	F	Pr > F
Tratamientos	3	5 882.357	1960.785	21.09	0.0001
Error	16	1 487.400	92.9625		
Total	19	7 369.757	—		

Donde se tiene evidencia para rechazar Ho, ya que Pr > F es mucho menor de 0.05.

DESPUÉS DEL ANÁLISIS DE VARIANZA

¿Cuál de todos los pares de medias son diferentes?

Para responder a esta pregunta se realizan pruebas de comparaciones múltiples de medias, como la de Tukey.

PRUEBA DE TUKEY

Este método se basa en utilizar el cuadrado medio del error, que se obtiene de un ANOVA. Para calcular un

valor ω que se compara con las diferencias de cada par de medias, si el resultado es mayor de ω se asumen medias diferentes en caso contrario se consideran semejantes o iguales.

La fórmula de cálculo es.

$$\omega = q_{\alpha,k,\nu}\sqrt{\frac{CM_{Erro}}{n}}$$

donde:
k = número de tratamientos o niveles
ν = grados de libertad asociados al CM_{Error}, con $\nu = N - k$
n = número de observaciones en cada uno de los k niveles
α = nivel de significación
$q_{\alpha,k,\nu}$ = valor crítico de rangos estudentizados (tablas)

La bibliografía reporta una amplia gama de pruebas, siendo las más comunes, además de la de Tukey, la de Fisher y la de Dunnet.

La prueba de Tukey y la de Fisher comparan todos los pares de medias, aunque Tukey genera intervalos más amplios que la de Fisher. Recomendando Tukey en estudios iniciales y la de Fisher en estudios finales o concluyentes.

La prueba de Dunnet permite comparar las medias contra un valor de referencia o control y dependiendo del "paquete" puede ser el primero o el último nivel del factor en estudio.

Después de comparar las medias, se recomienda verificar el cumplimiento de supuestos, para avalar la calidad de las conclusiones a las que se llega a través del análisis realizado: Homocedasticidad, Normalidad y comportamiento de residuales.
Homocedasticidad, varianzas homogéneas o significativamente iguales entre todos los tratamientos, aquí se recomienda la prueba de Barttlet

PRUEBA DE *BARTTLET* PARA HOMOGENEIDAD DE VARIANZAS

Esta prueba considera el siguiente par de hipótesis.

Ho: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, Todas las varianzas son iguales
Ha: Al menos dos varianzas son diferentes

Consiste básicamente en obtener un estadístico de contraste cuya distribución se aproxima a una distribución ji-cuadrada, con $k - 1$ grados de libertad, cuando las k muestras aleatorias son de poblaciones normales e independientes. La secuencia de cálculo es.

1. Considerando la fórmula

$$\chi^2 = 2.3026\frac{q}{c}$$

2. Obtener $s_p^2 = \dfrac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{N - k}$

3. Utilizar este resultado para calcular

$$q = (N-k)\log_{10} s_p^2 - \sum_{j=1}^{k}(n-1)\log_{10} s_j^2$$

4. Calcular

$$c = 1 + \frac{1}{3(k-1)}\left(\sum_{j=1}^{k}(n_j - 1)^{-1} - (N-k)^{-1}\right)$$

5. Obtener el valor calculado de ji-cuadrada y compararlo con el valor de tablas con nivel de significación α y $k - 1$ grados de libertad. Regla de decisión: Si $\chi^2_{calculada} > \chi^2_{k,\nu}$ se rechaza Ho.

PRUEBA DE *LEVENE* MODIFICADA

Debido a que la prueba de Bartlett es sensible al supuesto de normalidad, hay situaciones donde se recomienda un procedimiento alternativo, como lo es éste método robusto en cuanto a las desviaciones de la normalidad, ya que se basa en las medianas y no en las medias de los tratamientos. La secuencia de cálculo es:

Primero y antes que nada considerar el par de hipótesis a trabajar.
Ho: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ Todas las varianzas son iguales
Ha: Al menos dos varianzas son diferentes

1. Obtener la mediana de cada tratamiento: $\tilde{Y}_j$

2. Obtener el valor absoluto de la desviación de cada observación con respecto a la mediana de su tratamiento. $d_{ij} = \left|Y_{ij} - \tilde{Y}_j\right|$

3. Sobre la tabla de estas diferencias, realizar un ANOVA y aplicar la regla de decisión sobre el estadístico F para rechazar o no la Hipótesis nula.

PRUEBAS DE NORMALIDAD

Otra prueba consiste en verificar si los datos se comportan de acuerdo a una distribución normal, para lo cual existen pruebas numéricas y gráficas. Las numéricas básicamente plantean una curva normal teórica y mediante una prueba de falta de ajuste someten a prueba la hipótesis nula de que los datos se apegan a la distribución (Método de Kolmogorov-Smirnov, Anderson-Darling). Otro método es el gráfico, el cual es más utilizado por su impacto visual y lo fácil de su interpretación.

GRÁFICOS DE PROBABILIDAD NORMAL

Estos gráficos permiten juzgar hasta donde un conjunto de datos puede o no ser caracterizado por una distribución de probabilidad específica, en este caso la normal.

Gráficos de probabilidad acumulada.

i	Observación Y_i	Yi en orden ascendente	p_i (%)	z_i	q_i
1	9.63	9.34	2.5	1.96	1.99
2	9.86	9.51	7.5	1.44	1.49
3	10.20	9.63	12.5	1.15	1.13
4	10.48	9.69	17.5	0.94	0.95
5	9.82	9.75	22.5	0.76	0.77
6	10.07	9.82	27.5	0.60	0.56
7	10.39	9.86	32.5	0.46	0.44
8	10.03	9.89	37.5	0.32	0.35
9	9.34	9.96	42.5	0.19	0.14
10	10.26	9.98	47.5	0.06	0.08
11	9.89	10.03	52.5	0.06	0.07
12	10.67	10.07	57.5	0.19	0.19
13	9.69	10.13	62.5	0.32	0.37
14	10.15	10.15	67.5	0.46	0.43
15	10.32	10.20	72.5	0.66	0.58
16	9.98	10.26	77.5	0.76	0.76
17	9.51	10.32	82.5	0.94	0.94
18	10.13	10.39	87.5	1.15	1.15
19	9.96	10.48	92.5	1.44	1.42
20	9.75	10.67	97.5	1.96	1.98

$$p_i = \frac{100(i-0.5)}{n}$$

$$q_i = \frac{Y_i - \overline{Y}}{s_Y}$$

Un gráfico de los pares (Y_i , p_i) se espera que tenga una forma de S para asegurar una aproximación normal, aunque es más común hacer este gráfico en papel normal para obtener una línea recta.

Si todos los puntos de los datos aparecen aleatoriamente distribuidos a lo largo de la línea recta y si la línea pasa sobre o cercanamente a la intersección de la media de Y, el ajuste de los datos a la distribución normal se considera adecuado con el 50% de probabilidad.

Contrariamente, si los puntos aparecen con forma de S, la sugerencia es que los datos no se distribuyen normalmente.

Con la ayuda de una tabla de probabilidad normal, las probabilidades acumuladas, p_i pueden convertirse en sus correspondientes valores normales estandarizados z_i.

$$P(z \le z_i) = p_i$$

Si se conoce la media, la varianza y la variable Y_i, los datos muestreados se pueden estandarizar utilizando la transformación:

$$q_i = \frac{Y_i - \mu_Y}{\sigma_Y}$$

dado que la μ_Y y σ_Y generalmente no se conocen, se usa la ecuación:

$$q_i = \frac{Y_i - \overline{Y}}{s_Y}$$

A continuación se puede hacer un gráfico de los puntos (q_i , z_i) que sirve para juzgar la normalidad de un conjunto de datos.

Si se traza una gráfica con la misma escala para q_i y z_i, se espera que los puntos se distribuyan aleatoriamente a lo largo de una línea recta dibujada a 45°.

El manejo de estos diseños en Statgraphics se muestra mediante ejemplos.

Ejemplo 1. Un fabricante supone que existe diferencia en el contenido de calcio en lotes de materia prima que le son suministrados por su proveedor. Actualmente hay una gran cantidad de lotes en la bodega. Cinco de estos son elegidos aleatoriamente. Un químico realiza cinco pruebas sobre cada lote y obtiene los siguientes resultados.

Lote				
1	2	3	4	5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

¿Hay diferencia significativa en el contenido de calcio de un lote a otro?

El par de hipótesis a probar es:
Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
Ha: al menos un par de medias es diferente

La secuencia de análisis es:

0. Ingresar los datos, con dos columnas, una para lote y otra para calcio.

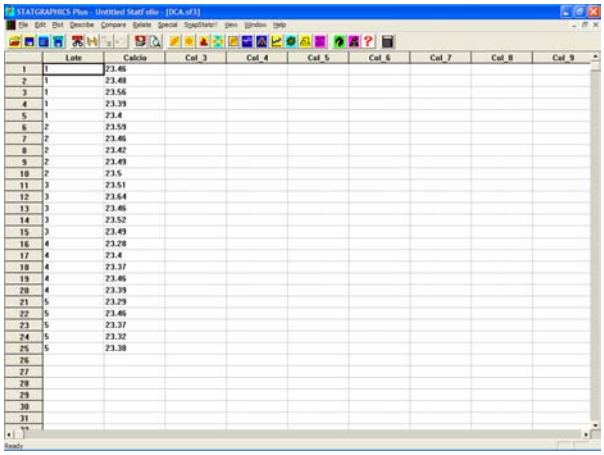


Figura 94. Hoja de Trabajo de STATGRAPHICS.

Seguir la secuencia:

1. **Compare→Analysis of Variance→One-Way ANOVA**

2. Colocar la variable **Calcio** en la caja **Dependent Variable**, ya que ésta es la respuesta o variable de interés y colocar la variable **Lote** en la caja **Factor**

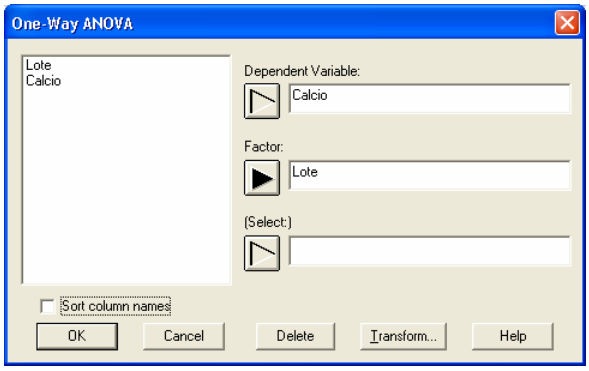


Figura 95. Diálogo para insertar variables.

3. En el diálogo de opciones tabulares, seleccionar todas las opciones, excepto **Table of Means** y **Kruskal Wallis Test**

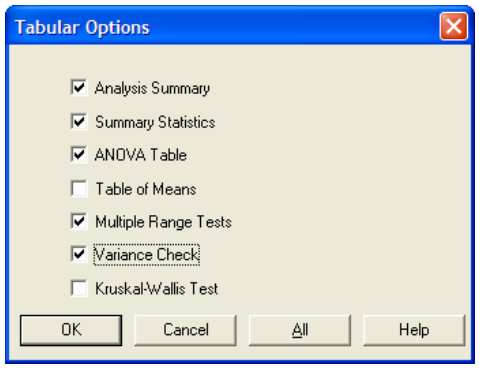


Figura 96. Opciones Tabulares.

La prueba de Kruskal-Wallis es una técnica No Paramétrica de Análisis de Varianza y se selecciona una vez que se comprueba la violación de supuestos del modelo del Diseño Completamente al Azar.

4. En el diálogo de opciones gráficas, seleccionar **Means Plot** y **todas las opciones de residuales.**

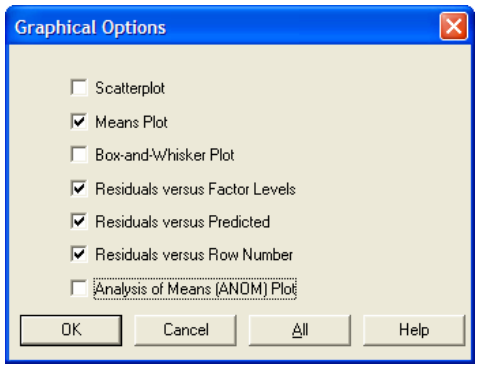


Figura 97. Opciones Grágicas.

Box-and-Whisker-Plot es una buena opción cuando se aplica la prueba de Kruskal-Wallis, ya que la muesca en los gráficos de cada tratamiento permite visualizar e interpretar la semejanza o diferencia entre tratamientos.

5. Dar un doble clic sobre cada una de las ventanas de resultados parciales y desactivar de las opciones tabulares o gráficas las que no sean de interés. Con un **clic derecho** y seleccionando **Copy Analysis to StatReporter**, se puede guardar en memoria todos los resultados presentes en pantalla.

6. Para guardar los resultados en disco, seguir la secuencia

File -> Save As -> Save StatReporter

En la caja de diálogo se le da nombre a un archivo en formato rtf (Rich Text File) que se puede "abrir" y trabajar en cualquier procesador de palabras, en nuestro caso se puede trabajar en WORD.

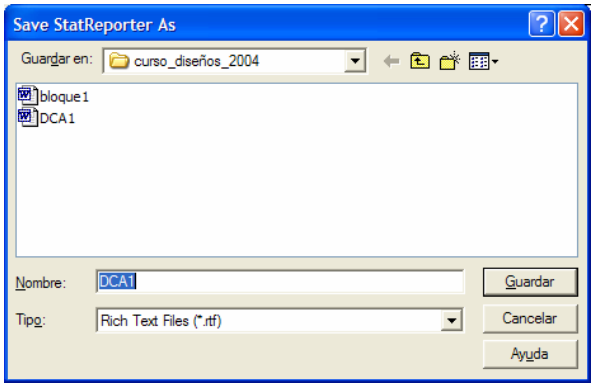


Figura 98. Guardar en disco.

RESULTADOS

One-Way ANOVA - Calcio by Lote

Analysis Summary

Dependent variable: Calcio
Factor: Lote
Number of observations: 25
Number of levels: 5

The StatAdvisor

This procedure performs a one-way analysis of variance for Calcio. It constructs various tests and graphs to compare the mean values of Calcio for the 5 different levels of Lote. The F-test in the ANOVA table will test whether there are any significant differences amongst the means. If there are, the Multiple Range Tests will tell you which means are significantly different from which others. If you are worried about the presence of outliers, choose the Kruskal-Wallis Test which compares medians instead of means. The various plots will help you judge the practical significance of the results, as well as allow you to look for possible violations of the assumptions underlying the analysis of variance.

En primer lugar se tiene una descripción de lo que hace el ONE-WAY ANOVA y posibles estrategias de análisis.

Summary Statistics for Calcio

Lote	Count	Average	Median	Mode	Variance
1	5	23.458	23.46		0.00472
2	5	23.492	23.49		0.00397
3	5	23.524	23.51		0.00473
4	5	23.38	23.39		0.00425
5	5	23.364	23.37		0.00423
Total	25	23.4436	23.46	23.46	0.00769067

Lote	Standard deviation	Minimum	Maximum	Range	Skewness
1	0.0687023	23.39	23.56	0.17	0.722536
2	0.0630079	23.42	23.59	0.17	0.892295
3	0.068775	23.46	23.64	0.18	1.60956
4	0.065192	23.28	23.46	0.18	-0.721849
5	0.0650385	23.29	23.46	0.17	0.603753
Total	0.0876964	23.28	23.64	0.36	0.137419

Lote	Kurtosis
1	-0.0789105
2	1.53018
3	3.13044
4	1.76747
5	0.294921
Total	0.0340129

The StatAdvisor

This table shows various statistics for Calcio for each of the 5 levels of Lote. The one-way analysis of variance is primarily intended to compare the means of the different levels, listed here under the Average column. Select Means Plot from the list of Graphical Options to display the means graphically.

Se tiene la estadística descriptiva por cada uno de los tratamientos (en este caso lotes)

ANOVA Table for Calcio by Lote

Analysis of Variance

Source	Sum of Squares	Df	Mean Squares	F-Ratio	P-Value
Between groups	0.096976	4	0.024244	5.54	0.0036
Within groups	0.0876	20	0.00438		
Total (Corr.)	0.184576	24			

The StatAdvisor

The ANOVA table decomposes the variance of Calcio into two components: a between-group component and a within-group component. The F-ratio, which in this case equals 5.53516, is a ratio of the between-group estimate to the within-group estimate. **Since the P-value of the F-test is less than 0.05, there is a statistically significant difference between the mean Calcio from one level of Lote to another at the 95.0% confidence level.** To determine which jeans are significantly different from which others, select Multiple Range Tests from the list of Tabular Options.

Hay que notar el texto en negritas, donde se indica que al menos un par de lotes son diferentes.

Multiple Range Tests for Calcio by Lote

Method: 95.0 percent Tukey HSD

Lote	Count	Mean	Homogeneous Groups
5	5	23.364	X
4	5	23.38	XX
1	5	23.458	XXX
2	5	23.492	XX
3	5	23.524	X

Contrast	Difference	+/- Limits
1 - 2	-0.034	0.12529
1 - 3	-0.066	0.12529
1 - 4	0.078	0.12529
1 - 5	0.094	0.12529
2 - 3	-0.032	0.12529
2 - 4	0.112	0.12529
2 - 5	*0.128	0.12529
3 - 4	*0.144	0.12529
3 - 5	*0.16	0.12529
4 - 5	0.016	0.12529

* denotes a statistically significant difference.

The StatAdvisor

This table applies a multiple comparison procedure to determine which means are significantly different from which others. The bottom half of the output shows the estimated difference between each pair of means. An asterisk has been placed next to 3 pairs, indicating that these pairs show statistically significant differences at the 95.0% confidence level. At the top of the page, 3 homogenous groups are identified using columns of X's. Within each column, the levels containing X's form a group of means within which there are no statistically significant differences. The method currently being used to discriminate among the means is Tukey's honestly significant difference (HSD) procedure. With this method, there is a 5.0% risk of calling one or more pairs significantly different when their actual difference equals 0.

Se tienen los grupos de medias semejantes, indicados por columnas de X's, así como por la matriz de comparaciones entre pares de medias. En este ejemplo se tienen 3 subgrupos, el primer subgrupo indica que los lotes 5, 4 y 1 son semejantes, el subgrupo 2 considera semejantes los lotes 4, 1 y 2 y por último son semejantes los lotes 1, 2 y 3.

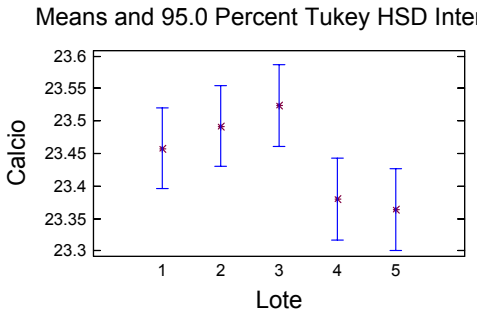
Variance Check

Cochran's C test: 0.215982 P-Value = 1.0
Bartlett's test: 1.00233 P-Value = 0.99978
Hartley's test: 1.19144
Levene's test: 0.0321932 P-Value = 0.997834

The StatAdvisor

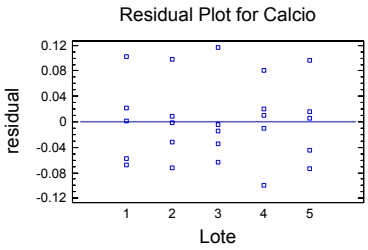
The four statistics displayed in this table test the null hypothesis that the standard deviations of Calcio within each of the 5 levels of Lote is the same. Of particular interest are the three P-values. Since the smallest of the P-values is greater than or equal to 0.05, there is not a statistically significant difference amongst the standard deviations at the 95.0% confidence level.

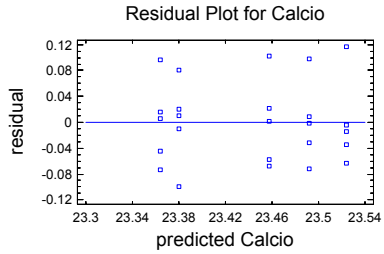
De las cuatro pruebas de homogeneidad de varianzas, las tres que muestran valores de probabilidad indican que se cumple el supuesto de homogeneidad de varianzas.



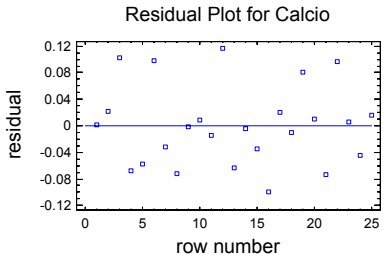
En este gráfico se puede apreciar que los lotes 1, 2 y 3 son semejantes entre ellos y los lotes 4 y 5 son semejantes entre ellos. También se aprecia que los lotes 1, 4 y 5 son semejantes, así como los lotes 2 y 4. Lo que se aprecia en los intervalos de confianza que muestran translape.

También se tiene que el lote 3 es el que presenta mayor concentración de calcio y el lote 5 el que tiene la más baja concentración.





Estos dos gráficos de residuales no muestran algún patrón que ponga en duda el cumplimiento del supuesto de normalidad y de homogeneidad de varianzas.



Este último gráfico no muestra tendencia alguna en los residuales, con respecto al número de hilera o fila, lo que demuestra que se cumple el supuesto de independencia.

Ejemplo 2. Tres diferentes soluciones para lavar están siendo comparadas con el objeto de estudiar su efectividad en el retraso del crecimiento de bacterias en envases de leche de 5 galones. El análisis se realiza en un laboratorio y sólo pueden efectuarse tres pruebas en un mismo día. Se hicieron conteos de colonias durante cuatro días. Analizar los datos y obtener conclusiones acerca de las soluciones.

	Días			
Solución	**1**	**2**	**3**	**4**
I	13	22	18	39
II	16	24	17	44
III	5	4	1	22

Este es un diseño de Bloques al azar, donde la variable de bloqueo es días y la variable a comparar es solución. Aquí ya no se puede realizar el análisis con el ONE-WAY ANOVA.

SOLUCIÓN

STATGRAPHICS cuenta con un módulo de Diseños Experimentales, al cual se accede al seleccionar del menú las opciones:

Special -> Experimental Design

Donde en esencia se hacen tres actividades: **Create** (definir el diseño), **Open** (ingresar datos), **Analyze** (hacer ANOVA).

Create, está opción despliega una caja de diálogo, donde se puede seleccionar el tipo de diseño experimental a realizar.

En esta caja de diálogo es importante asegurarse de tener un 1 en **No. Of Response Variables** (Número de variables de respuesta) y en **No. Of Experimental Factors** (Número de factores experimentales).

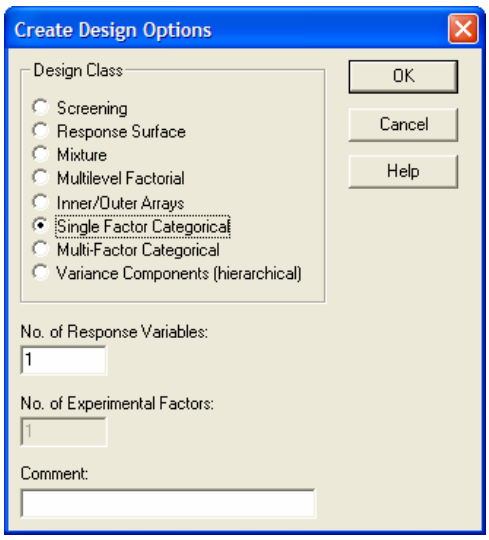


Figura 99. Creación del diseño.

Al dar OK aparece un diálogo para definir el factor en estudio, con su nombre, número de niveles y hasta las unidades en las que se mide.

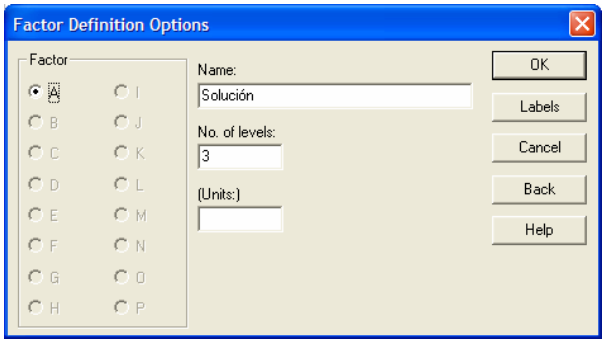


Figura 100. Definición del factor.

El siguiente diálogo permite definir las características de la respuesta: Nombre y unidades de medición

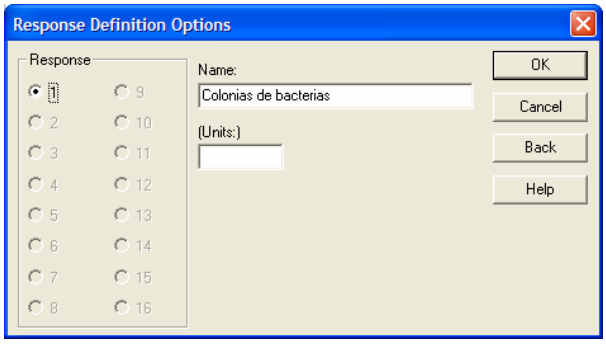


Figura 101. Definición de la variable respuesta.

A continuación aparece un diálogo donde se define el tipo de diseño a realizar.

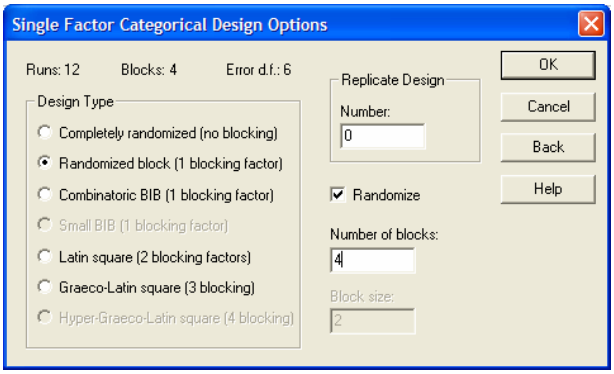


Figura 102. Elección del diseño.

Es importante notar la selección del Diseño de Bloques Aleatorizados (**Randomized block**), en número de réplicas o repeticiones de cero, ya que no se repite el experimento y 4 bloques.

La opción **Randomize** puede activarse o no. En la etapa de planeación experimental ayuda a establecer una secuencia en la que se debe realizar el experimento. Si ya se tienen datos, es más conveniente no activarla, para facilitar la captura de datos.

Es importante notar el número de "corridas" igual a 12.

Al dar OK, aparece una ventana de resultados que resume el tipo de diseño que se ha creado.

ingle Factor Categorical Design Attributes

```
Design Summary
--------------
Design class: Single Factor Categorical
File name: <Untitled>

Base Design
-----------
Number of experimental factors: 1
Number of blocks: 4
Number of responses: 1
Number of runs: 12
Error degrees of freedom: 6
Randomized: Yes

Factors   Levels  Units  Responses        Units
------------------------------------------------
Solución    3           Colonias de bacteria


The StatAdvisor
---------------
   You have created a Randomized block design
consisting of 12 runs. The design is to be run
in 4 blocks. The order of the experiments has
been fully randomized. This will provide
protection against the effects of lurking
variables.
```

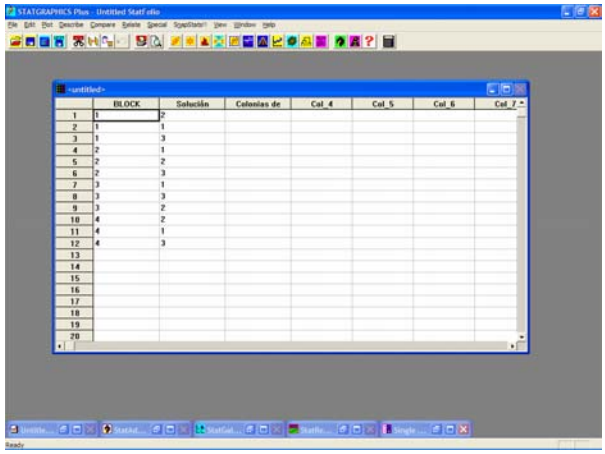


Figura 103. Diseño listo para guardarse.

Se genera una tabla de datos con los valores definidos para el factor y el número de bloques. Así como para capturar los datos resultantes del experimento a analizar.

Este archivo se puede guardar en disco, con la secuencia:

Save as -> Save Design File as

Este archivo se almacena con extensión SFX y sólo se puede abrir en el módulo de diseños de experimentos de Statgraphics.

Un diseño se abre para agregar datos o para su análisis con la secuencia.

Special -> Experimental Design -> Open Design

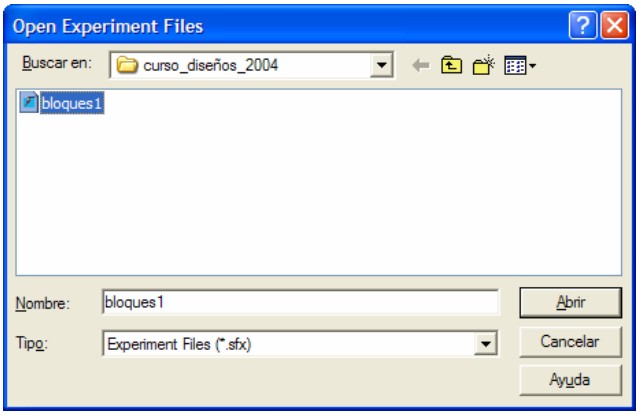


Figura 104. Abriendo el diseño para capturar las respuestas.

Para realizar el análisis, sobre el diseño que esté abierto en ese momento se sigue la secuencia.

Special -> Experimental Design -> Analyze Design

Para iniciar el análisis colocar la variable de respuesta en la caja **DATA** y presionar el botón **OK.**

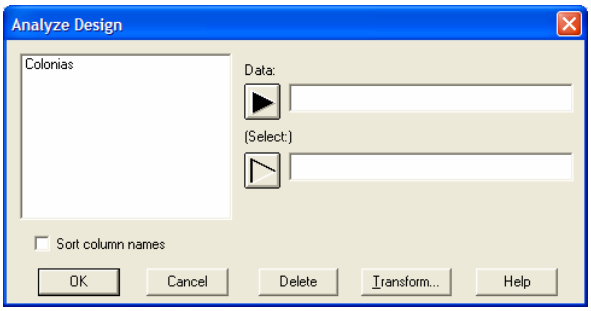


Figura 105. Realizando el análisis.

Al aparecer los resultados se pueden seleccionar las opciones Tabulares y Gráficas.

Seleccionar todas las opciones de la caja de diálogo **Tabular Options.**

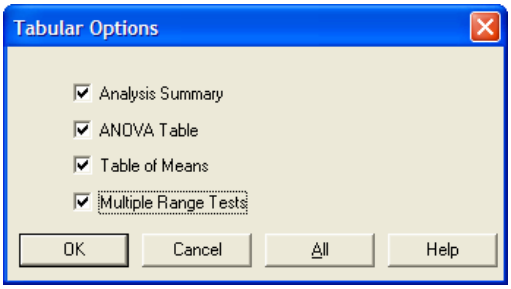


Figura 106. Opciones tabulares.

En las opciones gráficas seleccionar todas las opciones de residuales y el gráfico de medias.

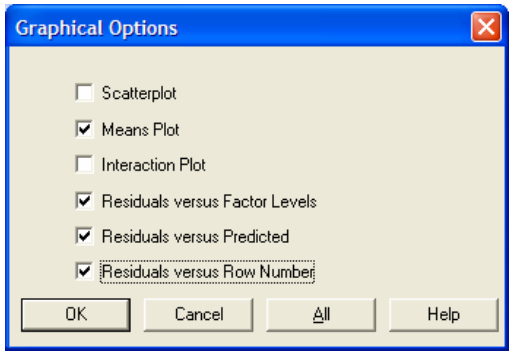


Figura 107. Opciones gráficas.

RESULTADOS

Multifactor ANOVA - Colonias

Analysis Summary

```
Dependent variable: Colonias
Factors:
        Solución
        BLOCK

Number of complete cases: 12
```

```
The StatAdvisor
---------------
   This procedure performs a multifactor analysis of variance for Colonias.  It constructs various tests and graphs to determine which factors have a statistically significant effect on Colonias.  It also tests for significant interactions amongst the factors, given sufficient data.  The F-tests in the ANOVA table will allow you to identify the significant factors.  For each significant factor, the Multiple Range Tests will tell you which means are significantly different from which others. The Means Plot and Interaction Plot will help you interpret the significant effects.  The Residual Plots will help you judge whether the assumptions underlying the analysis of variance are violated by the data.
```

Analysis of Variance for Colonias - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Solución	703.5	2	351.75	40.72	0.0003
B:BLOCK	1106.92	3	368.972	42.71	0.0002
RESIDUAL	51.8333	6	8.63889		
TOTAL (CORRECTED)	1862.25	11			

All F-ratios are based on the residual mean square error.

The StatAdvisor

The ANOVA table decomposes the variability of Colonias into contributions due to various factors. Since Type III sums of squares (the default) have been chosen, the contribution of each factor is measured having removed the effects of all other factors. **The P-values test the statistical significance of each of the factors. Since 2 P-values are less than 0.05, these factors have a statistically significant effect on Colonias at the 95.0% confidence level.**

Se tiene evidencia de que al menos un par de soluciones son diferentes. También se tiene evidencia del efecto de bloque, por lo es un acierto hacer bloques, aunque se debe recalcar que la presencia de una F en el bloque no implica que se deba de interpretar.

Table of Least Squares Means for Colonias with 95.0 Percent Confidence Intervals

Level	Count	Mean	Stnd. Error	Lower Limit	Upper Limit
GRAND MEAN	12	18.75			
Solución					
1	4	23.0	1.4696	19.404	26.596
2	4	25.25	1.4696	21.654	28.846
3	4	8.0	1.4696	4.40401	11.596
BLOCK					
1	3	11.333	1.69695	7.18104	15.485
2	3	16.666	1.69695	12.5144	20.819
3	3	12.0	1.69695	7.84771	16.152
4	3	35.0	1.69695	30.8477	39.152

The StatAdvisor

This table shows the mean Colonias for each level of the factors. It also shows the standard error of each mean, which is a measure of its sampling variability. The rightmost two columns show 95.0% confidence intervals for each of the means. You can display these means and intervals by selecting Means Plot from the list of Graphical Options.

En esta tabla se tiene una visión general de las medias, tanto para los tratamientos o niveles del factor, como para los bloques.

Multiple Range Tests for Colonias by Solución

Method: 95.0 percent Tukey HSD

Solución	Count	LSMean	LSSigma	Homogeneous groups
3	4	8.0	1.4696	X
1	4	23.0	1.4696	X
2	4	25.25	1.4696	X

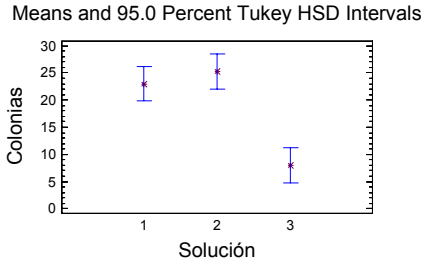
Contrast	Difference	+/- Limits
1 - 2	*-2.25	0.00000419388
1 - 3	*15.0	0.00000419388
2 - 3	*17.25	0.00000419388

* denotes a statistically significant difference

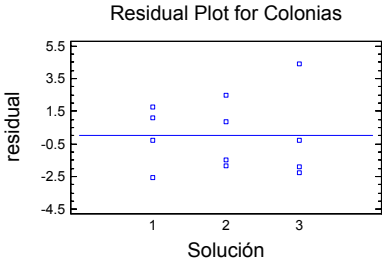
The StatAdvisor

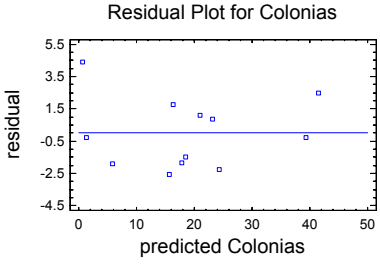
This table applies a multiple comparison procedure to determine which means are significantly different from which others. The bottom half of the output shows the estimated difference between each pair of means. An asterisk has been placed next to 3 pairs, indicating that these pairs show statistically significant differences at the 95.0% confidence level. At the top of the page, 3 homogenous groups are identified using columns of X's. Within each column, the levels containing X's form a group of means within which there are no statistically significant differences. The method currently being used to discriminate among the means is Tukey's honestly significant difference (HSD) procedure. With this method, there is a 5.0% risk of calling one or more pairs significantly different when their actual difference equals 0.

La prueba de Comparaciones de medias, muestra que la solución 3 es diferente de las soluciones 1 y 2, además de ser la que presenta una media menor.

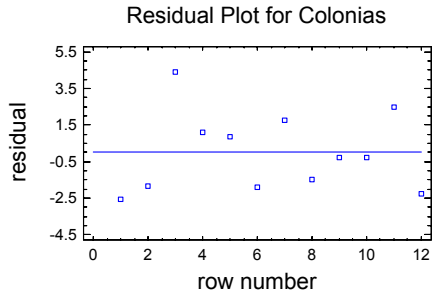


Se corrobora que la solución 3 es diferente de las otras 2 y que presenta el menor número de colonias.





No se hace patente algún patrón claro que muestre desviaciones de la normalidad y de la homogeneidad de la varianza. Aunque se aprecia un valor residual relativamente alto al extremo superior izquierdo de la gráfica.



La conclusión es que la media de la solución 3 es diferente a las otras dos y es la que menos colonias presenta en el conteo, lo que implica que es la mejor solución, esto a la luz de los datos.

DISEÑOS FACTORIALES

Este tipo de diseños permiten analizar varios factores a la vez, considerando su interacción.

La construcción típica de un factorial **axb** se presenta a continuación, donde **a** indica el número de niveles del primer factor y **b** el del segundo factor.

	FACTOR B				TOTAL
FACTOR A	1	2	…	*b*	$Y_{i..}$
1	Y_{111} Y_{112} Y_{113} Y_{114}	Y_{121} Y_{122} Y_{123} Y_{124}	…	Y_{1b1} Y_{1b2} Y_{1b3} Y_{1b4}	$Y_{1..}$
2	Y_{211} Y_{212} Y_{213} Y_{214}	Y_{221} Y_{222} Y_{223} Y_{224}	…	Y_{1b1} Y_{1b2} Y_{1b3} Y_{1b4}	$Y_{2..}$
. . .	. . .	. . .	. . .	. . .	. . .
a	Y_{a11} Y_{a12} Y_{a13} Y_{a14}	Y_{a21} Y_{a22} Y_{a23} Y_{a24}	…	Y_{ab1} Y_{ab2} Y_{ab3} Y_{ab4}	$Y_{a..}$
Total $Y_{.j.}$	$Y_{.1.}$	$Y_{.2.}$	…	$Y_{.b.}$	$Y...$

Modelo: $Y_{ijk} = \mu + \tau_i + \beta_j + \tau_i\beta_j + \varepsilon_{ijk}$

con: $i = 1,2,\cdots,a$; $j = 1,2,\cdots,b$ y $k = 1,2,\cdots,n$

1. *Ho:* $\tau_i = 0$ v.s. *Ha:* $\tau_i \neq 0$; para al menos una i.
2. *Ho:* $\beta_j = 0$ v.s. *Ha:* $\beta_j \neq 0$; para al menos una j.
3. *Ho:* $\tau_i\beta_j = 0$ v.s. *Ha:* $\tau_i\beta_j \neq 0$ para al menos un par $i \neq j$.

$$SC_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{...})^2$$

$$SC_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{i..} - \overline{Y}_{...})^2 = bn\sum_{i=1}^{a}(\overline{Y}_{i..} - \overline{Y}_{...})^2$$

$$SC_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{.j.} - \overline{Y}_{...})^2 = an\sum_{j=1}^{b}(\overline{Y}_{.j.} - \overline{Y}_{...})^2$$

$$SC_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2 =$$

$$= n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2$$

$$SC_{Error} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij.})^2$$

Grados de Libertad

$A = a - 1$
$B = b - 1$
$AB = (a-1)(b-1)$
$Error = ab(n-1)$
$Total = abn - 1 = N - 1$

¿TABLA DE ANALISIS DE VARIANZA, PARA UN DISEÑO: AxBxCxD?

Se puede ver a través de algunos ejemplos

Ejemplo 3. Se encuentra en estudio el rendimiento de un proceso químico. Se cree que las dos variables más importantes son la temperatura y la presión. Seleccionando para el estudio tres temperaturas y tres presiones diferentes, obteniendo los siguientes resultados de rendimiento.

	Presión		
Temperatura	**Baja**	**Media**	**Alta**
Baja	90.4 90.2	90.7 90.6	90.2 90.4
Intermedia	90.1 90.3	90.5 90.6	89.9 90.1
Alta	90.5 90.7	90.8 90.9	90.4 90.1

Aplicar el modelo adecuado y sacar las conclusiones pertinentes.

SOLUCIÓN

0. Hipótesis:

1. *Ho:* $\tau_i = 0$ (No existe efecto de temperatura) v.s.
   Ha: $\tau_i \neq 0$; para al menos una i (Existe efecto de temperatura).
2. *Ho:* $\beta_j = 0$ (No existe efecto de presión) v.s.
   Ha: $\beta_j \neq 0$; para al menos una j (Existe efecto de presión).
3. *Ho:* $\tau_i \beta_j = 0$ (No existe efecto de interacción) v.s.
   Ha: $\tau_i \beta_j \neq 0$ para al menos un par $i \neq j$ (Existe efecto de interacción).

1. Entrar al módulo de Diseños Experimentales, seleccionando del menú las opciones:

Special -> Experimental Design

2. En la opción **Create** se despliega una caja de diálogo, donde se puede seleccionar el tipo de diseño experimental a realizar.

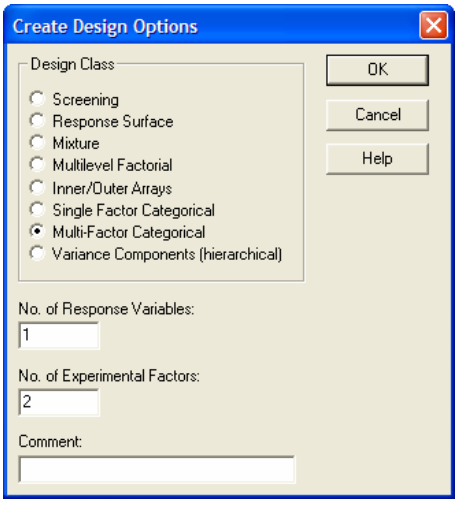


Figura 108. Creación del diseño.

3. En esta caja de diálogo es importante asegurarse de tener un 1 en **No. Of Response Variables** (Número de variables de respuesta) y un 2 en **No. Of Experimental Factors** (Número de factores experimentales).

En el diálogo que aparece se definen los factores en estudio, con su nombre, número de niveles y hasta las unidades en las que se mide.

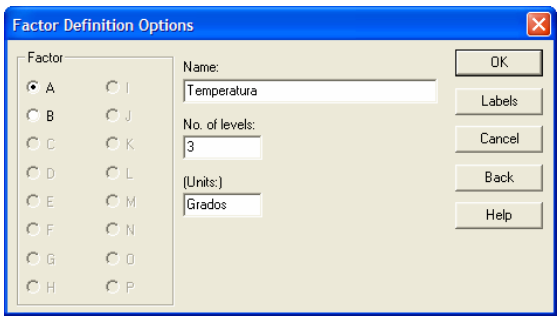


Figura109. Definición de los factores de estudio.

Hay que dar un clic en el identificador del factor (A o B) para acceder a su definición.

4. El siguiente diálogo permite definir las características de la respuesta: Nombre y unidades de medición

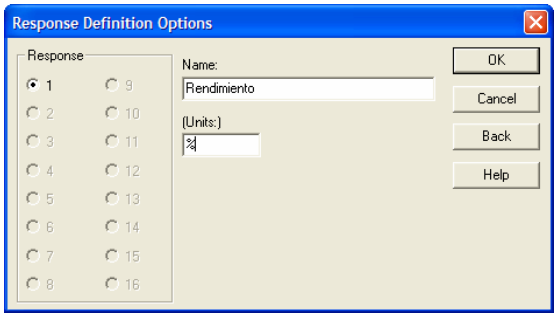


Figura 110. Definición de la variable respuesta.

5. A continuación aparece un diálogo para definir el número de replicas o repeticiones del experimento, así como si el diseño se debe generar de manera aleatoria.

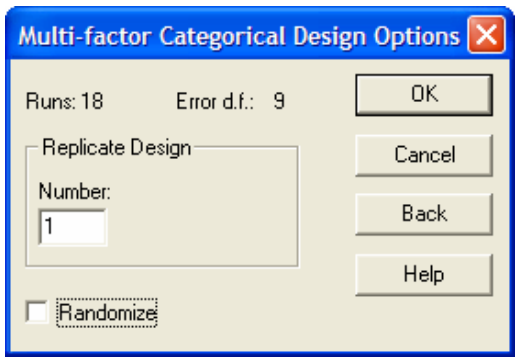


Figura 111. Número de réplicas.

La opción **Randomize** puede activarse o no. En la etapa de planeación experimental ayuda a estableces una secuencia en la que debe realizarse el experimento. Si ya se tienen datos, es más conveniente no activarla para facilitar la captura de datos.

Al dar OK, aparece una ventana de resultados que resume el tipo de diseño que se ha creado.

```
Multi-factor Categorical Design Attributes
Design Summary
--------------
Design class: Multi-factor Categorical
File name: <Untitled>

Base Design
-----------
Number of experimental factors: 2
Number of responses: 1
Number of runs: 18                Error degrees
of freedom: 9
Randomized: No

Factors                Levels      Units
-----------------------------------------------
Temperatura            3           Grados
Presión                3

Responses              Units
-----------------------------------
Rendimiento            %
```

6. En este momento es conveniente almacenar este diseño en disco, lo que se logra con la secuencia.

File -> Save As -> Save Design File As …

Este archivo se almacena con extensión SFX y sólo se puede abrir en el módulo de diseños de experimentos de Statgraphics.

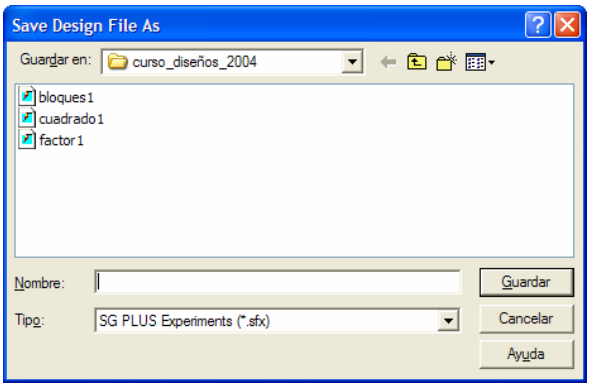


Figura112. Guardando el diseño.

7. Un diseño se abre para agregar datos o para su análisis con la secuencia.

Special -> Experimental Design -> Open Design

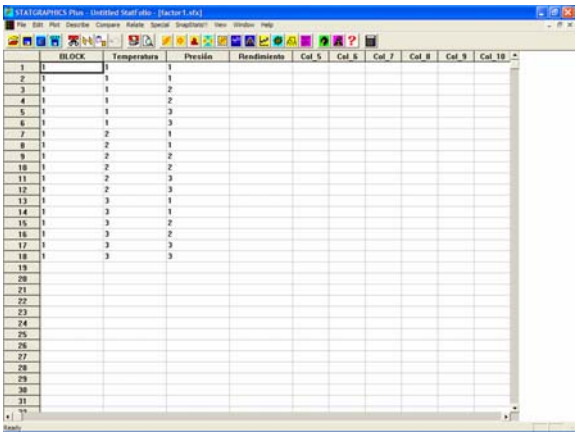


Figura113. Agregar datos.

8. Para realizar el análisis, sobre el diseño que esté abierto en ese momento se sigue la secuencia.

Special -> Experimental Design -> Analyze Design

Para iniciar el análisis colocar la variable de respuesta en la caja **DATA** y presionar el botón **OK.**

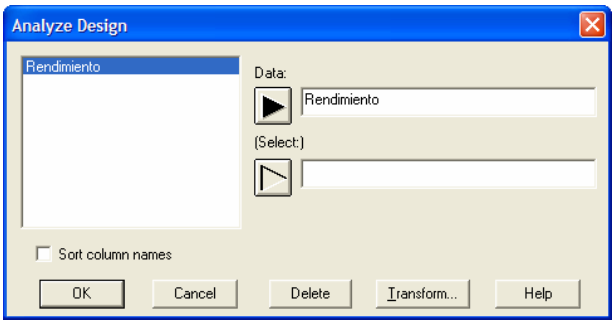


Figura114. Introdcir la variable respuesta.

9. Al aparecer los resultados se pueden seleccionar las opciones Tabulares y Gráficas.

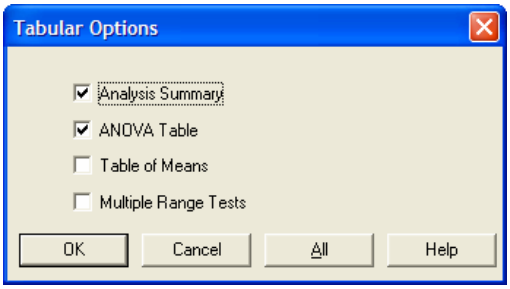


Figura115. Opciones tabulares.

El interés se centra en la Tabla de ANOVA de la caja de diálogo **Tabular Options.**

10. En las opciones gráficas seleccionar todas las opciones de residuales y el gráfico de interacción.

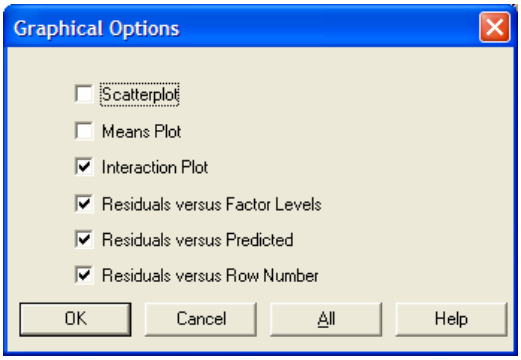


Figura116. Opciones Gráficas.

RESULTADOS
Multifactor ANOVA - Rendimiento

```
Analysis Summary

Dependent variable: Rendimiento
Factors:
        Temperatura
        Presión

Number of complete cases: 18


Analysis of Variance for Rendimiento - Type III
Sums of Squares
-----------------------------------------------
Source  Sum of  Df  Mean       F      P-Value
       Squares      Square
-----------------------------------------------
MAIN EFFECTS
A:Temperatura
      0.301111  2  0.150556    8.47    0.0085
B:Presión
      0.767778  2  0.383889    21.6    0.0004
INTERACTIONS
AB    0.068889  4  0.017222     0.97   0.4700

RESIDUAL  0.16  9  0.0177778
-----------------------------------------------
TOTAL (CORRECTED)
      1.29778  17
-----------------------------------------------
All F-ratios are based on the residual mean
square error.
```

En esta tabla se tiene evidencia del efecto de la presión y la temperatura (ya que los P-values son menores que 0.05 en ambos casos), aunque el efecto de interacción es no significativo.

NOTA: En algunos casos puede no presentarse el efecto de interacción, por lo que al dar un clic derecho sobre la ventana del ANOVA, seleccionar ANALYSIS OPTIONS.

En el diálogo que aparece indicar que se requieren las interacciones de orden 2.

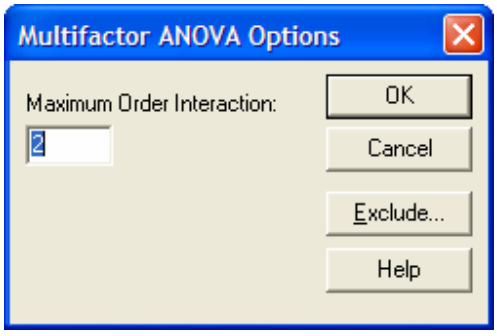
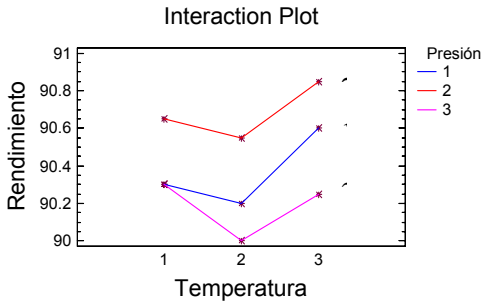


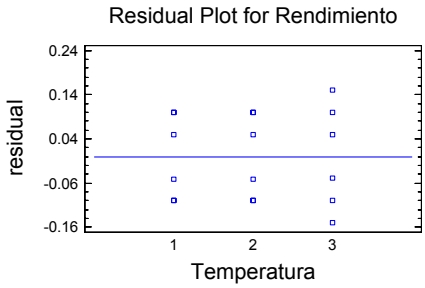
Figura117. Opciones de análisis.

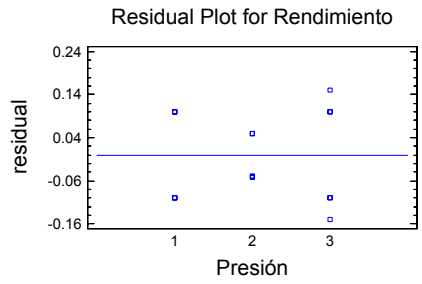
El gráfico de interacciones es el elemento clave en un diseño factorial, ya que es aquí donde se realiza la interpretación y se sustenta la toma de decisiones



Por ejemplo, aquí se puede indicar a que temperatura y a que presión se obtiene el máximo rendimiento. En este caso en la combinación de factores: temperatura alta y presión media.

El siguiente paso es analizar el cumplimiento de supuestos.





Donde los gráficos de residuales muestran que la temperatura 3 y la presión 3 presentan mayor variabilidad que los otros 2 niveles.

En estos gráficos no se aprecia algún patrón claro lo que permite asumir que si se están cumpliendo los supuestos, dándole validez a las conclusiones del experimento.

Se puede "jugar" con **Pane Options** para cambiar la variable a incluir en el gráfico.

Ejemplo 2. En una operación de lotes se produce un químico viscoso, donde cada lote produce suficiente producto para llenar 100 contenedores. El ensayo del producto es determinado por análisis infrarrojo que realiza duplicado alguno de los 20 analistas del laboratorio. En un esfuerzo por mejorar la calidad del producto se realizó un estudio para determinar cual de tres posibles fuentes de variabilidad eran significativas en el proceso y su magnitud.

Las fuentes seleccionadas fueron: la variable A lotes, se seleccionaron aleatoriamente tres lotes de producción mensual, la variable analistas, B, seleccionando dos de manera aleatoria, la variable C corresponde a dos contenedores seleccionados de manera aleatoria de cada lote. Obteniendo los siguientes resultados.

Lote	No. de Contenedor			
	I		II	
	Analista		Analista	
	M	P	M	P
23	94.6 95.2	95.8 95.8	97.7 98.1	97.8 98.6
35	96.2 96.4	96.5 96.9	98.0 98.4	99.0 99.0
2	97.9 98.1	98.4 98.6	99.2 99.4	99.6 100.0

SOLUCIÓN

Diseño factorial 3x2x2, lo que interesa es la significancia de los efectos principales, así como la de cada una de las dobles interacciones y la triple interacción.

1. *Ho:* $\tau_i = 0$ (No existe efecto de lote) v.s.
   Ha: $\tau_i \neq 0$; para al menos una i (Existe efecto de lote).
2. *Ho:* $\beta_j = 0$ (No existe efecto de contenedor) v.s.
   Ha: $\beta_j \neq 0$; para al menos una j (Existe efecto de contenedor).
3. *Ho:* $\gamma_l = 0$ (No existe efecto de analista) v.s.
   Ha: $\gamma_l \neq 0$; para al menos una l (Existe efecto de analista).
4. *Ho:* $\tau_i\beta_j = 0$ (No existe efecto de interacción de lote y contenedor) v.s.
   Ha: $\tau_i\beta_j \neq 0$ para al menos un par $i \neq j$ (Existe efecto de interacción de lote y contenedor).
5. *Ho:* $\tau_i\gamma_l = 0$ (No existe efecto de interacción de lote y analista) v.s.
   Ha: $\tau_i\gamma_l \neq 0$ para al menos un par $i \neq l$ (Existe efecto de interacción de lote y analista).
6. *Ho:* $\beta_j\gamma_l = 0$ (No existe efecto de interacción de contenedor y analista) v.s.
   Ha: $\beta_j\gamma_l \neq 0$ para al menos un par $j \neq l$ (Existe efecto de interacción de contenedor y analista).
7. *Ho:* $\tau_i\beta_j\gamma_l = 0$ (No existe efecto de interacción de lote, contenedor y analista) v.s.
   Ha: $\tau_i\beta_j\gamma_l \neq 0$ para al menos un par $i \neq j \neq l$ (Existe efecto de interacción de lote, contenedor y analista).

1. Acceder al módulo de Diseños Experimentales, seleccionando del menú las opciones:

   Special -> Experimental Design

2. En la opción **Create** se despliega una caja de diálogo, donde se puede seleccionar el tipo de diseño experimental a realizar.

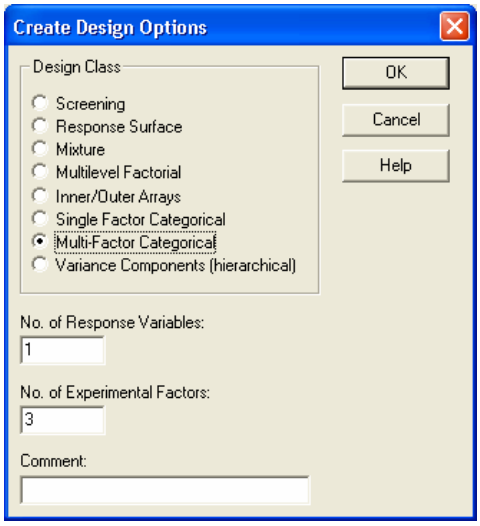


Figura118. Crear el diseño

En esta caja de diálogo es importante asegurarse de tener un 1 en **No. Of Response Variables** (Número de variables de respuesta) y un 3 en **No. Of Experimental Factors** (Número de factores experimentales).

A continuación todo el proceso es análogo al ejemplo anterior, recalcando la importancia de analizar los gráficos de interacciones.

NOTA: Otra opción es teclear directamente los datos y del menú seguir la secuencia.

Compare -> Analysis Of Variance -> Multifactorial Anova

EJERCICIOS

1. Se encuentra bajo estudio el efecto que tienen 5 reactivos distintos (A, B, C, D y E) sobre el tiempo de reacción de un proceso químico. Cada lote de material nuevo es lo suficientemente grande para permitir que sólo se realicen 5 ensayos. Más aún, cada ensayo tarda aproximadamente una hora y media por lo que sólo pueden realizarse cinco ensayos por día. En el experimento se busca controlar sistemáticamente las variables lote de material y día, ¿que se puede decir del tiempo de reacción de los 5 reactivos diferentes?

	Día				
Lote	**1**	**2**	**3**	**4**	**5**
1	A,8	B,7	D,1	C,7	E,3
2	C,11	E,2	A,7	D,3	B,8
3	B,4	A,9	C,10	E,6	D,5
4	D,6	C,8	E,6	B,1	A,10
5	E,4	D,2	B,3	A,8	C,8

Recomendación: Diseño de Cuadrados Latinos, con día y lote como variables de bloqueo. Se hace igual que el ejemplo 2, pero en el modelo también se incluye la variable lote.

2. En un experimento para comparar el porcentaje de eficiencia de cuatro diferentes resinas quelantes (A, B, C y D) en la extracción de iones de Cu^{2+} de solución acuosa, el experimentador sólo puede realizar cuatro corridas con cada resina. De manera que durante tres días seguidos se preparó una solución fresca de iones Cu^{2+} y se realizó la extracción con cada una de las resinas, tomadas de manera aleatoria, obteniendo los siguientes resultados. ¿Cuál es el modelo más adecuado para analizar este experimento y cuales son sus conclusiones?

	Resinas			
Día	**A**	**B**	**C**	**D**
1	97	93	96	92
2	90	92	95	90
3	96	91	93	91
4	95	93	94	90

Recomendación: Diseño de Bloques al azar, con día como variable de bloqueo.

3. Se llevó a cabo un experimento para probar los efectos de un fertilizante nitrogenado en la producción de lechuga. Se aplicaron cinco dosis diferentes de nitrato de amonio a cuatro parcelas (réplicas). Los datos son el número de lechugas cosechadas de la parcela.

Tratamiento (Kg N/Ha)				
0	104	114	90	140
50	134	130	144	174
100	146	142	152	156
150	147	160	160	163
200	131	148	154	163

Recomendación: Diseño Completamente al Azar (One-Way).

CAPÍTULO 9

ANÁLISIS DE REGRESIÓN

Problemas que se plantean:

1) ¿Cuál es el modelo matemático más apropiado para describir la relación entre una o más variables independientes $(X_1, X_2, \cdots, X_k)$ y una variable dependiente (Y)?

2) Dado un modelo especifico, ¿qué significa éste y cómo se encuentran los parámetros del modelo que mejor ajustan a nuestros datos? Si el modelo es una línea recta: ¿cómo se encuentra la "mejor recta"?

La ecuación de una línea recta es:

$$Y = f(X) = \beta_0 + \beta_1 X$$

Donde:

β_0 ordenada al origen

β_1 pendiente

En un análisis de regresión lineal simple, el problema es encontrar los valores que mejor estimen a los parámetros β_0 y β_1. A partir de una muestra aleatoria.

El modelo de regresión lineal es:

$$Y_i = \mu_{Y/X} + \varepsilon_i = \beta_0 + \beta_1 X + \varepsilon_i \text{, con } i = 1,2,3,\cdots,n$$

Para cada observación el modelo es:

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$
$$\vdots$$
$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

Que se puede escribir como:

$$_nY_1 = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad _nX_2 = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ & \vdots \\ 1 & X_n \end{pmatrix} \quad _2\beta_1 = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad _n\varepsilon_1 = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

donde:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = Y = X\beta + \varepsilon$$

Estimación por mínimos cuadrados

Sea $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ la respuesta estimada en X_i con base en la línea de regresión ajustada. La distancia vertical entre el punto (X_i, Y_i) y el punto $(X_i, \hat{Y}_i)$ de la recta ajustada está dada por el valor absoluto de $\left|Y_i - \hat{Y}_i\right|$ o $\left|\hat{Y}_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right|$, cuya suma de cuadrados es:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

El problema ahora es encontrar los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que $\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$ sea mínimo.

Solución:

Si Q = $\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$, entonces

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \ = 0 \qquad (1)$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)\ (-X_i) = 0 \qquad (2)$$

(NOTA: las derivadas parciales se igualan a cero para determinar los puntos críticos, que serán mínimos). Esto conduce a las **Ecuaciones Normales de Mínimos Cuadrados**

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

En notación matricial se tiene:

$$\begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^{\,2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

$$\mathbf{X^t X \beta = X^t Y}$$

de donde

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{Y})$$

Solución matricial para calcular los parámetros de la ecuación de regresión

La solución algebraica de las ecuaciones normales, para datos muestrales, genera las siguientes ecuaciones:

$$b_0 = \frac{\sum_{i=}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$b_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

ALGO DE GEOMETRÍA

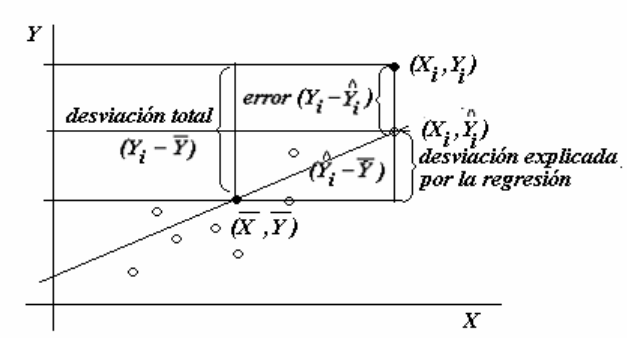


Figura119. Desviaciones: total, explicada por la regresión y error.

1) $Y_i - \overline{Y}$ desviación total

2) $\hat{Y}_i - \overline{Y}$ desviación explicada por la regresión

3) $Y_i - \hat{Y}_i$ error

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

Total = Regresion + Error

Al aplicar sumatorias y elevar al cuadrado se tiene:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}[(\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)]^2$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

$SC_{Total} = SC_{Regresion} + SC_{Error}$

Cantidades que permiten realizar un ANOVA, para contrastar las hipótesis:

$$Ho: \beta_i = 0$$
$$Ha: \beta_i \neq 0$$

Fuente de variación	g.l.	SC	CM	F_C	F_t
Regresión	1	SC_{Reg}	CM_{Reg}	$\frac{CM_{Reg}}{CM_{Error}}$	$F_{1-\alpha,1,n-2}$
Error Residual	$n-2$	SC_{Error}	CM_{Error}		
Total	$n-1$	SC_{Total}			

Este ANOVA considera el siguiente par de hipótesis:

Ho: $\beta_i = 0$, es decir que todos los coeficientes del modelo son iguales a cero y por lo tanto no hay un modelo lineal que describa el comportamiento de los datos.

Contra *Ha*: $\beta_i \neq 0$ de que al menos uno de los coeficientes es diferente de cero y entonces si hay un modelo lineal.

INTERPRETANDO a β_0 y β_1

Ho: $\beta_1 = 0$

Caso 1.- *Ho*: $\beta_1 = 0$ **No se rechaza. Es decir que la pendiente es cero o que no hay pendiente,** entonces se tienen dos opciones de interpretación.

a) Si la suposición de línea recta es correcta significa que X no proporciona ayuda para predecir Y, esto quiere decir que $\overline{Y}$ predice a Y.

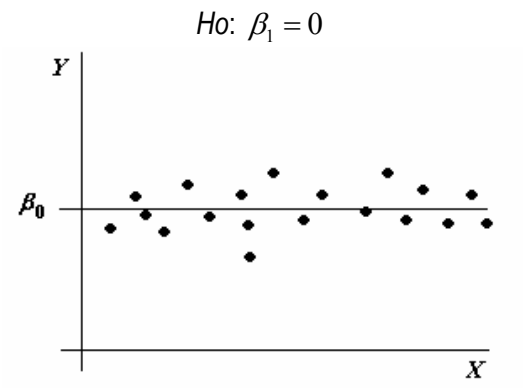


Figura120. *X* no proporciona ayuda para predecir *Y*.

b) La verdadera relación entre X y Y no es lineal, esto significa que el modelo puede involucrar funciones cuadráticas, cúbicas o funciones más complejas.

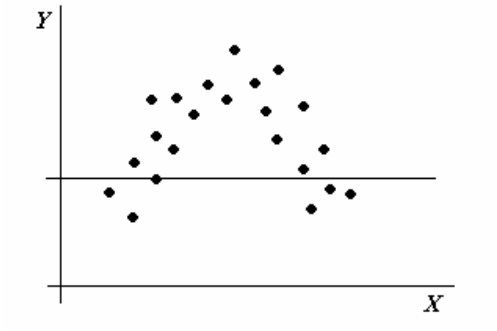


Figura121. La relación entre X y Y no es lineal.

NOTA: Si hay una curvatura se requiere un elemento cuadrático en el modelo, si hay dos curvaturas entonces se requiere un cúbico y así sucesivamente.

Caso 2.- *Ho*: $\beta_1 = 0$ **se rechaza (es decir, si hay pendiente o en otras palabras si hay un modelo lineal que describe el comportamiento de los datos).**

a) X proporciona información significativa para predecir Y

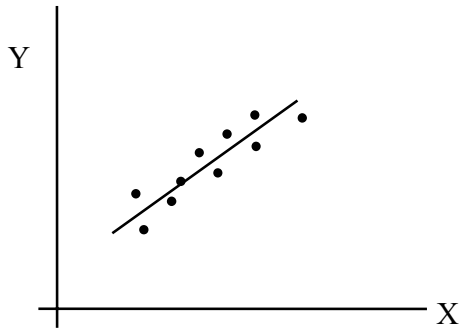


Figura122. La relación entre X y Y es lineal.

b). El modelo puede tener un término lineal más, quizás un término cuadrático.

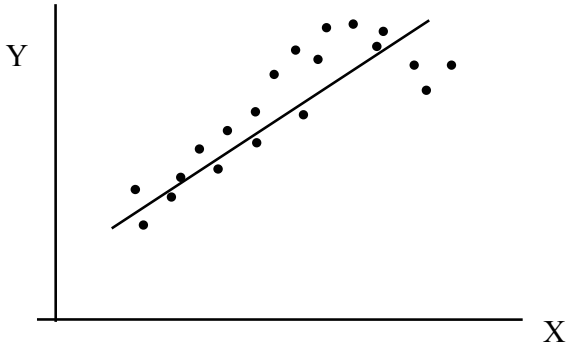


Figura123. La relación entre X y Y no es lineal.

Caso 3. Prueba. *Ho*: $\beta_0 = 0$, Si NO se rechaza esta Hipótesis, puede ser apropiado ajustar un modelo sin β_0, siempre y cuando exista experiencia previa o teoría que sugiera que la recta ajustada debe pasar por el origen y que existan datos alrededor del origen para mejorar la información sobre β_0.

CORRELACION

Si X y Y son dos variables aleatorias (no existe causa-efecto), entonces el coeficiente de correlación se define como:

1) r $\in$ [-1,1]

2) r es independiente de las unidades de X y Y

3) $\hat{\beta}_1 > 0 \Leftrightarrow r > 0$

$\hat{\beta}_1 < 0 \Leftrightarrow r < 0$

$\hat{\beta}_1 = 0 \Leftrightarrow r = 0$

r es una medida de la fuerza de asociación lineal entre X y Y

NOTA: NO se puede ni se deben establecer relaciones causales a partir de los valores de r, ya que ambas variables son aleatorias.

COEFICIENTE DE DETERMINACIÓN r^2

$$r^2 = \frac{SC_{total} - SC_{error}}{SC_{total}} = \frac{SC_{Reg}}{SC_{Total}}$$

donde, $r^2 \in [0,1]$

Esta r-cuadrada es una medida de la variación de Y explicada por los cambios o variación en la X. Es común leerla como porcentaje de variación en Y explicada por los cambios en X.

DIAGNÓSTICO DEL MODELO DE REGRESIÓN LINEAL SIMPLE

Las técnicas de diagnóstico son esenciales para detectar desacuerdos entre el modelo y los datos para los cuales se ajusta éste. Esto se hace a través del análisis de los residuos.

Los supuestos que se hacen del estudio del análisis de regresión son:

- La relación entre *Y* y *X* es lineal.
- Los errores tienen media cero
- Los errores tienen varianza constante σ^2.
- Los errores no están correlacionados (son independientes).
- Los errores se distribuyen normalmente.

Las posibles violaciones al modelo se pueden detectar a través de los residuos y son:

- Evidencias que sugieren que la forma del modelo no es la apropiada.
- Presencia de casos extraordinarios (outliers) en los datos.
- Evidencia que sugieren varianza no constante.
- Evidencia de que la distribución de los errores no proviene de una distribución normal.
- Autocorrelación, que se define como la falta de independencia de los residuos (errores).

REGRESION NO-LINEAL

En ocasiones, la relación $Y - X$ presenta una tendencia curvilínea, entonces se debe recurrir a los ajustes no lineales. En estos casos es importante tener una idea más o menos clara del tipo de curva al que se debe ajustar, ya que hay: logarítmicas, cuadráticas, cúbicas, inversas, potenciales, logísticas o exponenciales, entre otras opciones.

REGRESION LINEAL MULTIPLE

Esta regresión se refiere a modelos lineales cuando se consideran dos o más variables independientes.

$$Y = f(X_1, X_2, \cdots, X_k)$$

Comparando la regresión simple contra la múltiple se tiene que:

1) Es más difícil la elección del mejor modelo, ya que casi siempre hay varias opciones razonables.

2) Se dificulta visualizar el modelo, por la dificultad de presentar más de tres dimensiones.

3) Requiere cálculos complejos, generalmente se realiza con recursos computacionales y software especializado.

4) Además de los supuestos arriba mencionados para la regresión lineal simple, en la regresión lineal múltiple se debe cumplir la no colinealiadad de las variables explicativas.

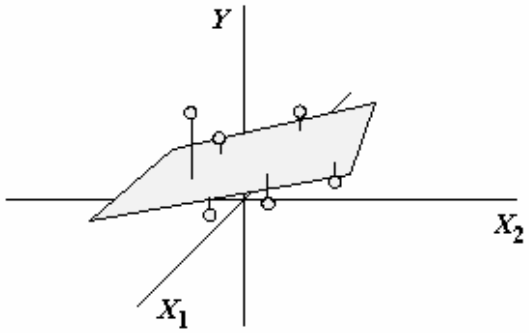


Figura124. Ajuste de un plano lineal con dos variables independientes

MÍNIMOS CUADRADOS

Al igual que en la regresión lineal simple, se puede trabajar el método de mínimos cuadrados**.** Para esto:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

donde:

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

Con base en los datos muestrales

$$Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k)$$

Al elemento de la derecha se le conoce como residual y refleja la desviación de los datos observados con respecto al plano ajustado.

Suma de cuadrados, elevando al cuadrado y sumando los elementos de la ecuación anterior.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k) \right]^2$$

El método consiste en encontrar los valores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots$ llamados estimadores de mínimos cuadrados, para los cuales la suma de cuadrados es mínima. De tal manera que, se pueda construir la siguiente tabla de ANOVA.

Tabla de ANOVA para las hipótesis:
Ho: $\beta_i = 0$
Ha: Al menos un $\beta_i \neq 0$

Fuente de variación	***g.l.***	***SC***	***CM***	***F_C***	***r^2***
Regresión	k	$SC_{Tot\text{-}Error}$	$\frac{SC_{\mathbf{Reg}}}{k}$	$\frac{CM_{\mathbf{Reg}}}{CM_{Error}}$	$\frac{SC_{Reg}}{SC_{tot}}$
Error o Residual	$n-k-1$	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$\frac{SC_{Error}}{n-k-1}$		
Total	$n-1$	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Donde los supuestos del análisis de Regresión se pueden resumir en la siguiente expresión.

$$\varepsilon \sim \mathrm{NI}(\mu_{Y/X_1,X_2,\cdots,X_k}, \sigma^2)$$

Los errores o residuales se distribuyen normal e independientemente con desviaciones al ajuste lineal igual a cero y varianza σ^2.

CORRELACIÓN PARCIAL y PARCIAL MÚLTIPLE

Medida de la fuerza de relación lineal entre dos variables, después de controlar los efectos de otras variables en el modelo.

Cuya representación está dada por:

$$R_{Y,X_1/X_2} \qquad R_{Y,X_1/X_2,X_3} \qquad R_{Y,(X_3,X_4,X_5)/X_1,X_2}$$

Expresiones que se leen:

$R_{Y,X_1/X_2}$ Correlación de las variables Y–X_1, cuando se tiene controlado el efecto de X_2 en un modelo. También se puede leer: correlación de Y–X_1, cuando X_2 ya está en el modelo.

$R_{Y,X_1/X_2,X_3}$ Correlación de las variables Y–X_1, cuando se tienen controlados los efectos de X_2 y X_3 en un modelo.

$R_{Y,(X_3,X_4,X_5)/X_1,X_2}$ Correlación de las variables X_3, X_4 y X_5 con Y, cuando se tienen controlados los efectos de X_1 y X_2 en un modelo.

CORRELACION Y DETERMINACIÓN MÚLTIPLE

$$R_{y/(x_1,x_2,\ldots,x_k)} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(\hat{Y}_i - \overline{\hat{Y}})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}})^2}}$$

$$R^2_{Y/(X_1,X_2,\cdots,X_k)} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

$$= \frac{SC_{total} - SC_{Error}}{SC_{Total}}$$

Donde r y r^2 representan la correlación y determinación simple, mientras que R y R^2 se utilizan para la correlación y determinación múltiple.

F's PARCIALES

La F's parciales son una herramienta útil para verificar si el ingreso o eliminación de una variable o grupos de variable mejoran el ajuste de un modelo lineal.

Este método de verificación se inicia con algunas preguntas, suponiendo 3 variables X_1, X_2 y X_3

1) ¿Se puede predecir el valor de Y utilizando sólo X_1?

2) ¿Adicionar X_2 contribuye significativamente en la predicción de Y, una vez que se considera la contribución de X_1?

3) ¿Contribuye X_3, dados X_1 y X_2 en el modelo?

Las respuestas a estas preguntas se obtienen al contrastar las siguientes hipótesis:

Ho: La adición de X* al modelo, incluyendo $X_1, X_2,..., X_k$, no mejora significativamente la predicción de Y.

Ho: $\beta^* = 0$, donde β^* es el coeficiente de X^*, en la ecuación de regresión.
Ha: $\beta^* \neq 0$

Cuyo estadístico de prueba es:

$$t_c = \frac{\hat{\beta}^*}{s_{\beta^*}}$$

Cuya regla de decisión es: rechazar Ho si $t_c > t_{1-\frac{\alpha}{2},n-k-1}$

ASPECTOS PRÁCTICOS DE LA REGRESIÓN LINEAL MÚLTIPLE

Para determinar la relación entre dos o más variables de regresión $X_1, X_2, \cdots, X_k$ y la respuesta Y. El problema general consiste en ajustar el modelo.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

Usualmente los parámetros desconocidos (β_k) se denominan coeficientes de regresión y pueden determinarse mediante **mínimos cuadrados**. Donde ε denominado error aleatorio debe presentar una media igual a cero y su varianza σ^2 no debe estar correlacionada.

Pruebas de hipótesis de la regresión lineal múltiple

A menudo se desea probar que tan significantes son los parámetros del modelo de regresión, lo cual se logra al contrastar si dichos coeficientes son iguales a cero; las hipótesis son:

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0$$

$$H_a : \beta_i \neq 0$$

Rechazar *Ho* implica que al menos una de variables del modelo contribuye significativamente al ajuste. El parámetro para probar esta hipótesis es una generalización del utilizado en regresión lineal simple. La suma total de cuadrados (SC_{Total}) se descompone en la suma de cuadrados de regresión (SC_{Reg}) y en la sumas de cuadrados del error (SC_{Error}).

$$SC_{Total} = SC_{\mathrm{Re}g} + SC_{Error}$$

Consecuentemente el valor de *F* estimado se obtiene de la ecuación:

$$F_C = \frac{\frac{SC_{\mathrm{Re}g}}{k}}{\frac{SC_{Error}}{n-k-1}} = \frac{CM_{\mathrm{Re}g}}{CM_{Error}}$$

Valor que se compara con una $F_{1-\frac{\alpha}{2},k,n-k-1}$ de tablas. La regla de decisión es: Rechazar *Ho* si $F_C > F$ de tablas.

Criterio para la selección de variables

Es importante probar las hipótesis con respecto a los coeficientes de regresión individuales; tales pruebas son útiles para evaluar cada variable de regresión en el modelo. En ocasiones el modelo puede ser más efectivo si se le introducen variables adicionales o, quizá si se desechan una o más variables que se encuentran en el mismo.

Introducir variables al modelo de regresión provoca que la suma de cuadrados de la regresión aumente y que la del error disminuya. Se debe decidir si el incremento de la suma de cuadrados de la regresión es suficiente para garantizar el uso de la variable adicional en el modelo. Además si se agrega una variable poco importante al modelo se puede aumentar el cuadrado medio del error, disminuyendo así la utilidad del mismo.

La hipótesis para probar la significancia de cualquier coeficiente individual, por ejemplo β_i son:

$$Ho: \beta_i = 0$$
$$Ha: \beta_i \neq 0$$

Y la estadística apropiada para probar la ecuación es:

$$t_c = \frac{\hat{\beta}_i}{s_{\beta_i}}$$

Donde β_i es el coeficiente a contrastar y $s_{\beta i}$ es el error estándar del coeficiente a contrastar. La regla de decisión es: rechazar *Ho* si $|t_C| > t_{1-\alpha/2,n-k-1}$

Coeficiente de determinación R^2 y R^2 *ajustado*

Después de encontrar la recta de regresión, se debe de investigar que tan bien se ajusta el modelo a los datos mediante el cálculo de R^2.

Este coeficiente se construye con base en dos cantidades. La primera es la suma de los cuadrados minimizada denominada suma de cuadrados del error (SC_{Error}), la cual representa la suma de las desviaciones al cuadrado de los datos a la recta que mejor se ajusta. La segunda cantidad es la suma de cuadrados alrededor de la media $\bar{Y}$, y se conoce como la suma de cuadrados totales ($SC_{Tot.}$).

El valor de R^2 se define de la siguiente forma:

$$R^2 = \frac{SC_{Tot} - SC_{Error}}{SC_{Tot}} = 1 - \frac{SC_{\mathrm{Re}g}}{SC_{Tot}} = 1 - \frac{SC_{Error}}{SC_{Tot}}$$

Y se interpreta como el porcentaje de la suma de cuadrados total que es explicada por la relación lineal. ***Conviene aclarar que a pesar que R^2 es un buen indicador de la calidad del ajuste de regresión, no se debe usar como un criterio único de selección del modelo.***

Al agregar variables, a un modelo lineal, el coeficiente de correlación y de determinación siempre aumentan. Por lo que es importante no tomar como único criterio de selección de modelos el valor de R o R^2. Es mejor utilizar el coeficiente de determinación ajustado, que considera el número de variables independientes $X_1, X_2, \cdots, X_k$ en el modelo, y cuya fórmula de cálculo es.

$$R^2_{ajustada} = R^2_a = 1 - \frac{\frac{SC_{Error}}{n-k}}{\frac{SC_{Tot}}{n-1}} = 1 - \frac{(n-1)SC_{Error}}{(n-k)SC_{Tot}}$$

El criterio de selección de variables para un ajuste lineal es que la R^2 sea mayor y que el cuadrado medio del error sea más pequeño. De tal manera que al comparar a dos o más modelos el mejor es aquel con R^2 mayor y menor CM_{Error} .

MÉTODOS DE SELECCIÓN DE VARIABLES

Para realizar un ajuste lineal múltiple se tienen tres métodos clásicos de selección de variables.

- FORWARD, se basa en incluir variables al modelo, en función de su significación, evitando que entren las no significativas.
- BACKWARD En un principio incluye todas las variables al modelo y empieza a descartar las menos significativas, hasta quedarse únicamente con las significativas.
- STEPWISE combina los dos métodos anteriores para incluir y eliminar variables hasta quedarse con las más significativas.

DESPUÉS DEL ANÁLISIS DE REGRESIÓN

Hay que verificar supuestos, así como cuidar problemas de **multicolinealidad y autocorrelación**. También se puede realizar una prueba de falta de ajuste, la cual contrasta la *Ho* de que el modelo se ajusta y describe los datos muestrales (requiere tener repeticiones de Y para cada uno de los valores de X).

La multicolinealidad se define como la dependencia entre las variables explicativas y la **autocorrelación** como la falta de independencia de los residuos.

EJEMPLOS

Ejemplo 1

Relación de gastos médicos mensuales en relación con el tamaño de familia **(Regresión Lineal Simple).**

TAMAÑO DE FAMILIA	GASTOS MEDICOS MENSUALES (en dólares)
2	20
2	28
4	52
5	50
7	78
3	35
8	102
10	88
5	51
2	22
3	29
5	49
2	25

¿Existe evidencia para establecer una relación lineal entre el tamaño de la familia y los gastos médicos? ¿Si la respuesta es afirmativa, cual es la ecuación de esta relación? ¿Se cumplen los supuestos del análisis de regresión?

Solución

1. Crear el archivo de datos con dos columnas, una para la variable independiente (X) y otra para la variable dependiente (Y).

2. Del menú seguir la secuencia

Relate -> Simple Regresión

3. En el diálogo que aparece colocar en su lugar la variable dependiente (Y) y la independiente (X).

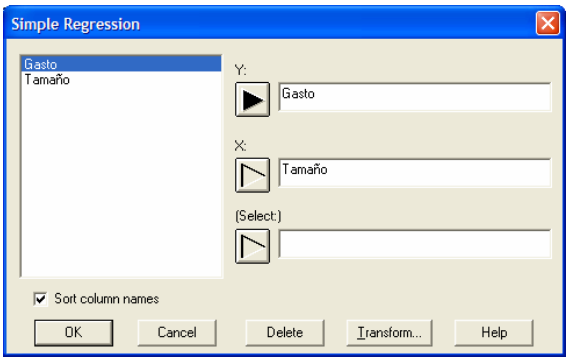


Figura125. Introducción de variables en el modelo de Regresión lineal simple.

4. Dar OK y llamar a las opciones tabulares y gráficas de esta opción.

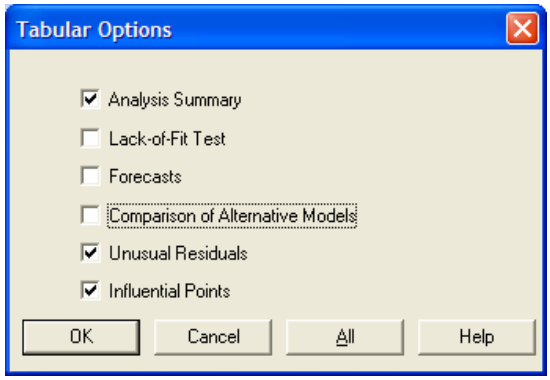


Figura126. Opciones Tabulares.

- **Forecast**s sirve para generar un modelo de regresión que se uitlice para hacer pronósticos.

- **Lack-of-Fit Test**, es una prueba para probar la falta de ajuste en el modelo. Funciona solamente cuando se tienen repeticiones de la variable Y en cada uno de los valores de X

- **Comparision of Alternative Models**. Con base en las correlaciones se define si hay algún otro modelo que explique mejor el comportamiento de los datos

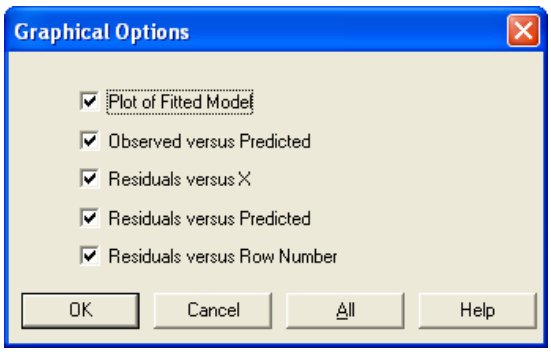


Figura127. Opciones Gráficas.

En las opciones gráficas se recomienda seleccionar todos los gráficos de residuales, ya que permiten analizar rápidamente el cumplimiento de supuestos.

RESULTADOS
Simple Regression - Gasto vs. Tamaño

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------
Dependent variable: Gasto
Independent variable: Tamaño
-----------------------------------------------
                     Stand.      T
Parameter  Estimate   Error  Statistic   P-Value
-----------------------------------------------
Intercept  4.70485   4.78913   0.982403   0.3470
Slope      9.79029   0.939226 10.4238     0.0000
-----------------------------------------------
```

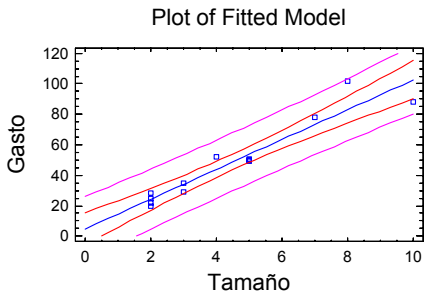
En esta primera tabla se prueba la hipótesis nula de que cada parámetro es cero.

Entonces la pendiente es diferente de cero (P-value = 0.0000 < 0.05 y se rechaza Ho); con la ordenada al origen se tiene evidencia estadística para considerarla cero. Esto significa que se puede hacer un ajuste considerando que la recta pasa por el origen.

```
                Analysis of Variance
-----------------------------------------------
Source     Sum of    Df   Mean     F-Ratio  P-Value
           Squares        Square
-----------------------------------------------
Model     7594.25    1  7594.25    108.66   0.0000
Residual  768.823 11     69.893
-----------------------------------------------
Total     8363.08    12
(Corr.)
```

Este ANOVA prueba la Hipótesis nula de que todos los coeficientes del modelo son iguales a cero. Al rechazarla se tiene evidencia estadística de que al menos un coeficiente es diferente de cero (ya se vió en la tabla anterior que es la pendiente)

```
Correlation Coefficient = 0.952927
R-squared = 90.8069 percent
R-squared (adjusted for d.f.) = 89.9712 percent
Standard Error of Est. = 8.3602
Mean absolute error = 5.72666
Durbin-Watson statistic = 2.5251 (P=0.1323)
Lag 1 residual autocorrelation = -0.274824

The StatAdvisor
---------------
   The output shows the results of fitting a
linear model to describe the relationship
between Gasto and Tamaño. The equation of the
fitted model is

   Gasto = 4.70485 + 9.79029*Tamaño

Since the P-value in the ANOVA table is less
than 0.01, there is a statistically significant
relationship between Gasto and Tamaño at the 99%
confidence level.

   The R-Squared statistic indicates that the
model as fitted explains 90.8069% of the
variability in Gasto. The correlation
coefficient equals 0.952927, indicating a
relatively strong relationship between the
variables. The standard error of the estimate
shows the Standard deviation of the residuals to
be 8.3602. This value can be used to construct
prediction limits for new observations by
selecting the Forecasts option from the text
menu.

   The mean absolute error (MAE) of 5.72666 is
the average value of the residuals. The Durban
Watson (DW) statistic tests the residuals to
determine if there is any significant
correlation based on the order in which they
occur in your data file. Since the P-value is
greater than 0.05, there is no indication of
serial autocorrelation in the residuals.
```



Esta gráfica muestra que si hay una tendencia lineal en los datos. En los resultados de correlación y determinación se tiene que modelo explica un 91% de la variación en los valores de Y, por efecto de los cambios en X.

La r^2 *ajustada* indica un 89.9712% de variación explicada de la variable Y por efecto de los cambios en la variable X. ANOVA para *Ho*: todos los coeficientes del modelo tienen valor cero vs la *Ha*: al menos uno de los coeficientes del modelo es diferente de cero.

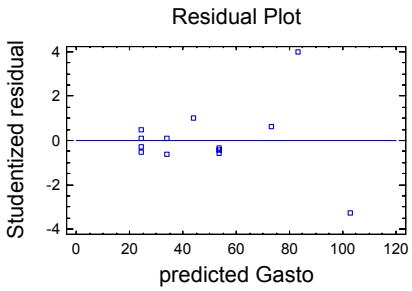
Modelo Gastos médicos mensuales = 4.70485 + 9.79029*(Tamaño de la familia)

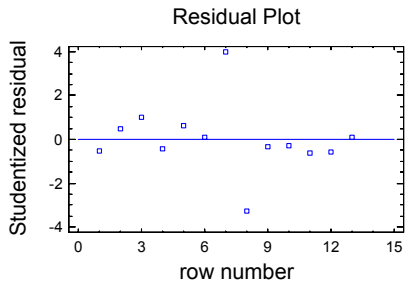
```
Comparison of Alternative Models
-----------------------------------------------
Model                Correlation       R-Squared
-----------------------------------------------
Multiplicative         0.9698           94.06%
Double reciprocal      0.9590           91.96%
Square root-X          0.9547           91.14%
Square root-Y          0.9536           90.93%
Linear                 0.9529           90.81%
S-curve               -0.9472           89.72%
Logarithmic-X          0.9410           88.54%
Exponential            0.9402           88.39%
Reciprocal-X          -0.8817           77.75%
Reciprocal-Y          -0.8780           77.08%
Logistic                      <no fit>
Log probit                    <no fit>
-----------------------------------------------
```

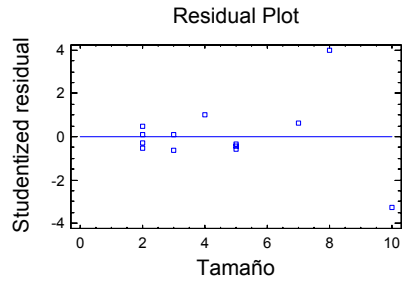
```
The StatAdvisor
---------------
   This table shows the results of fitting several curvilinear models to the data. Of the models fitted, the multiplicative model yields the highest R-Squared value with 94.0567%. This is 3.2498% higher than the currently selected linear model. To change models, select the Analysis Options dialog box.
```

Esta tabla muestra que los posibles modelos y su correlación para sugerir si hay un modelo que se ajuste mejor a los datos.

```
Unusual Residuals
-------------------------------------------------
                    Predicted           Studentized
Row    X     Y        Y          Residual    Residual
-------------------------------------------------
7     8.0   102.0    83.0272     18.9728      3.97
8    10.0    88.0   102.608     -14.6078     -3.28
-------------------------------------------------

The StatAdvisor
---------------
   The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Gasto deviates from a model fitted using all of the data except that observation.    In this case, there are 2 Studentized residuals greater than 3.0.  You should take a careful look at the observations greater than 3.0 to determine whether they are outliers which should be removed from the model and handled separately.

Influential Points
-------------------------------------------------
                 Predicted  Studentized
Row    X     Y       Y       Residual    Leverage
-------------------------------------------------
8    10.0   88.0   102.608    -3.28       0.464078
-------------------------------------------------
Average leverage of single data point = 0.153846

The StatAdvisor
---------------
   The table of influential data points lists all observations which have leverage values greater than 3 times that of an average data point.  Leverage is a statistic which measures how influential each observation is indetermining the coefficients of the estimated model. In this case, an average data point would have a leverage value equal to 0.153846.  There is one data point with more than 3 times the average leverage, but none with more than 5 times.
```







INTERPRETACIÓN

En resumen se puede decir:

- El análisis de varianza muestra que al menos uno de los coeficientes del modelo es diferente de cero, en otras palabras, el modelo es significativo, entonces si hay modelo.

- Después se tiene que la pendiente es diferente de cero, pero se tiene evidencia estadística de que la ordenada al origen se puede considerar cero. Entonces, se puede ajustar un modelo con ordenada igual a cero.

- Los gráficos de residuales muestra que si bien los datos no son completamente normales, tampoco tienen mucha desviación de la normalidad, por lo que las conclusiones son confiables, desde el punto de vista estadístico.

- Las propuestas de modelo indican que puede tener un mejor ajuste el modelo multiplicativo (potencial). Para probarlo sólo basta con dar un clic derecho en cualquier ventana de resultados, después un clic sobre **Analysis**

Options y seleccionar un tipo de modelo en el diálogo que aparece.

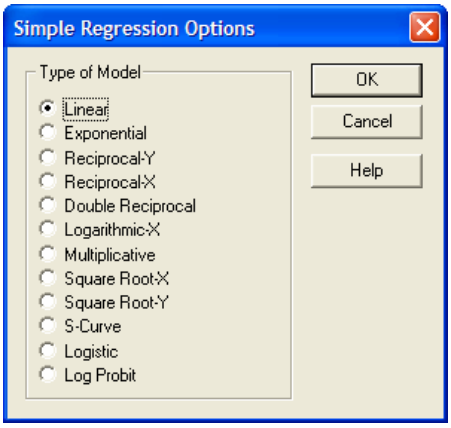


Figura128. Opciones para seleccionar diferentes modelos.

Ejemplo 2

El artículo "Determination of Biological Maturity and Effects of Harvesting and Drying Conditions of Milling Quality of Paddy" (J. Agricultural Eng. Research, 1975, pp. 353-361) reporta los siguientes datos sobre la fecha X de cosecha (número de días después de la floración) y producción Y (Kg/Ha) de arroz producido en la India (**Regresión No-Lineal). (Jay L. Devore, Probabilidad y estadística para ingeniería y ciencias, 5ª. Edición, Internacional Thomson Editores, 2001, pág. 555)**

X	16	18	20	22	24	26	28	30
Y	2508	2518	3304	3423	3057	3190	3500	3883

X	32	34	36	38	40	42	44	46
Y	3823	3646	3708	3333	3517	3241	3103	2776

SOLUCIÓN

1. Crear el archivo de datos con dos columnas, una para la variable independiente (X) y otra para la variable dependiente (Y).

2. Seguir la secuencia

Relate -> Polynomial Regresión

3. Llenar el diálogo que aparece, ubicando en su lugar la variable dependiente y la independiente.

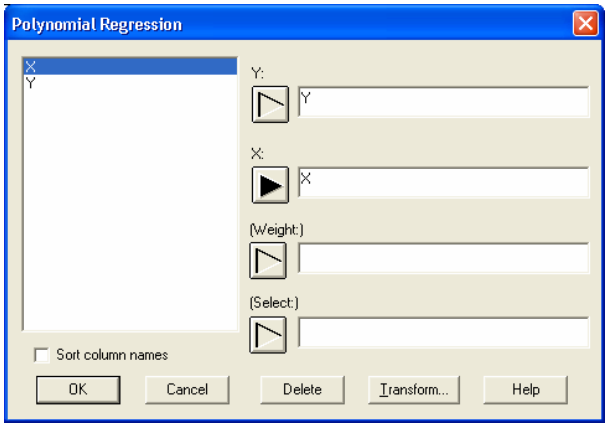


Figura129. Regresión polinómica.

4. Dar OK y seleccionar las Opciones Tabulares y Gráficas, de manera semejante a la Regresión Simple.

Resultados
Polynomial Regression - Y versus X

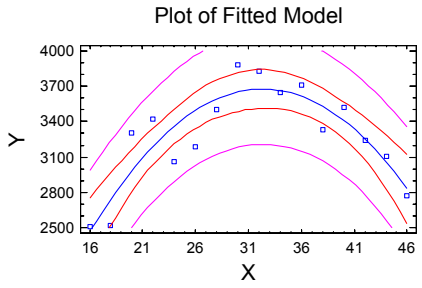
```
Polynomial Regression Analysis
----------------------------------------------
Dependent variable: Y
----------------------------------------------
                     Standard      T     P-Value
Parameter  Estimate    Error     Statistic
----------------------------------------------
CONSTANT -1070.4    617.253    -1.73413   0.1065
X          293.483   42.1776     6.95826   0.0000
X^2         -4.5358   0.674415 -6.72554    0.0000
----------------------------------------------

                Analysis of Variance
----------------------------------------------
Source    Sum of     Df   Mean    F-Ratio   P-Value
          Squares         Square
----------------------------------------------
Model     2.08478E6  2  1.04239E6  25.08    0.0000
Residual 540388.0   13  41568.3
----------------------------------------------
Total     2.62517E6 15
(Corr.)


R-squared = 79.4151 percent
R-squared (adjusted for d.f.) = 76.2482 percent
Standard Error of Est. = 203.883
Mean absolute error = 151.371
Durbin-Watson statistic = 1.96525 (P=0.2394)
Lag 1 residual autocorrelation = 0.0126911

The StatAdvisor
---------------
   The output shows the results of fitting a
second order polynomial model to describe the
relationship between Y and X.  The equation of
the fitted model is

Y = -1070.4 + 293.483*X-4.5358*X^2

Since the P-value in the ANOVA table is less
than 0.01, there is a statistically significant
relationship between Y and X at the 99%
confidence level.
```

The R-Squared statistic indicates that the model as fitted explains 79.4151% of the variability in Y. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 76.2482%.

The standard error of the estimate shows the standard deviation of the residuals to be 203.883. This value can be used to construct prediction limits for new observations by selecting the Forecasts option from the text menu.

The mean absolute error (MAE) of 151.371 is the average value of the residuals. The Durbin Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is greater than 0.05, there is no indication of serial autocorrelation in the residuals.

In determining whether the order of the polynomial is appropriate, note first that the P-value on the highest order term of the polynomial equals 0.0000141423. Since the P value is less than 0.01, the highest order term is statistically significant at the 99% confidence level. Consequently, you probably don't want to consider any model of lower order.



95.0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	-1070.4	617.253	-2403.89	263.099
X	293.483	42.1776	202.364	384.602
X^2	-4.5358	0.674415	-5.99279	-3.07881

The StatAdvisor

This table shows 95.0% confidence intervals for the coefficients in the model. Confidence intervals show how precisely the coefficients can be estimated given the amount of available data and the noise which is present.

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Y deviates from a model fitted using all of the data except that observation. In this case, there are no Studentized residuals greater than 2.0.

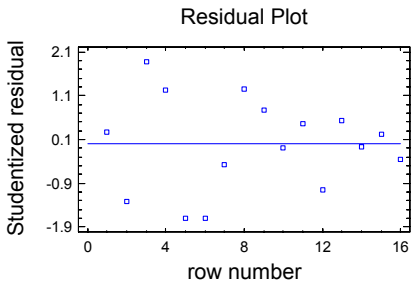
Influential Points

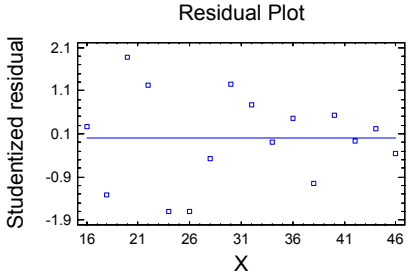
Row	Leverage	Mahalanobis Distance	DFITS

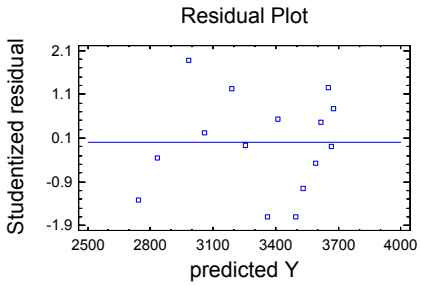
Average leverage of single data point = 0.1875

The StatAdvisor

The table of influential data points lists all observations which have leverage values greater than 3 times that of an average data point, or which have an unusually large value of DFITS. Leverage is a statistic which measures how influential each observation is in determining the coefficients of the estimated model. DFITS is a statistic which measures how much the estimated coefficients would change if each observation was removed from the data set. In this case, an average data point would have a leverage value equal to 0.1875. There are no data points with more than 3 times the average leverage. There are no data points with unusually large values of DFITS.







INTERPRETACIÓN

En el gráfico se aprecia que el modelo lineal no es la mejor opción de ajuste, ya que se presenta una curvatura, entonces se propone un modelo cuadrático. Modelo que se ajusta por "default" u omisión, aunque con un clic derecho aparece un menú flotante donde se puede seleccionar Opciones de Análisis y cambiar el orden a 3 = cúbico o uno de mayor orden.

Es importante considerar el valor de R^2 y de R^2 ajustado, este último con un valor del 76.24%

El ANOVA muestra que al menos uno de los coeficientes del modelo es diferente de cero, entonces si hay modelo. El modelo es Y = -1070.3976 + 293.4829X – 4.5358X^2, aunque existe evidencia de que la ordenada puede pasar por el origen, punto (0,0).

El siguiente paso es verificar los supuestos del modelo estadístico, que mediante los gráficos de residuales muestran que no hay desviaciones importantes de la normalidad, de la homogeneidad de varianzas ni de la independencia.

Por último se podría probar el modelo con un ajuste por el origen.

EJEMPLO 3.
(Regresión Lineal Múltiple)

Y	X_1	X_2	X_3
506	10	6	55
811	18	10	32
816	20	11	34
752	16	9	48
610	15	5	58
903	21	12	29
685	11	7	52
830	18	10	36
650	14	8	60
793	15	6	49
961	19	8	24
692	18	10	63
752	12	7	45
488	10	7	61
848	17	8	38
611	15	9	59
709	14	10	41
919	22	10	26
827	20	9	39
526	9	6	65

¿Hay relación lineal entre (X_1 , X_2 , X_3) con Y?

SOLUCIÓN

1. Ingresar los datos en 4 columnas, una para cada variable: Y, X_1, X_2 y X_3.

2. Seguir la secuencia

Relate -> Multiple Regression

3. En el diálogo que aparece colocar en el sitio correspondiente la variable de respuesta, Y, así como las independientes, X_1, X_2 y X_3.

4. Dar OK y en la ventana de resultados seleccionar opciones tabulares y gráficas.

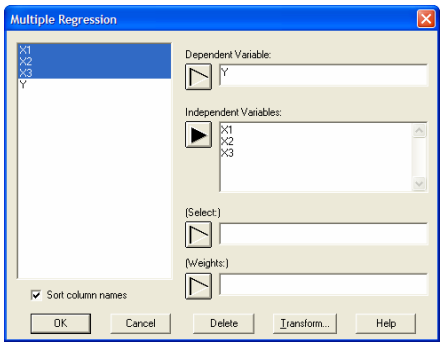


Figura130. Regresión lineal mútiple.

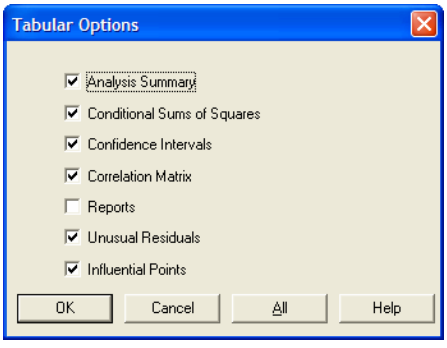


Figura131. Opciones Tabulares.

Considerar todas las opciones, menos la de **Reports**.

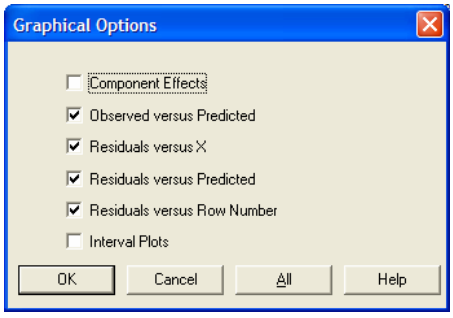


Figura132. Opciones Gráficas.

Opciones gráficas para verificar supuestos, sleccionar todas **excepto Component Effects e Inerval Plots**.

RESULTADOS

Multiple Regression - Y

```
Multiple Regression Analysis
-----------------------------------------------
Dependent variable: Y
-----------------------------------------------
                     Standard     T
Parameter   Estimate  Error     Statistic P-Value
-----------------------------------------------
CONSTANT    837.202  127.237    6.57987    0.0000
X1           17.4761   5.40612 3.23265    0.0052
X2           -9.96119  9.04575 -1.1012    0.2871
X3           -6.42127  1.33028 -4.82702   0.0002
-----------------------------------------------

                Analysis of Variance
-----------------------------------------------
Source    Sum of    Df   Mean      F-Ratio  P-Value
          Squares        Square
-----------------------------------------------
Model    319121.0    3   106374.0   41.56   0.0000
Residual 40947.9   16    2559.24
-----------------------------------------------
Total    360069.0   19
(Corr.)

R-squared = 88.6278 percent
R-squared (adjusted for d.f.) = 86.4955 percent
Standard Error of Est. = 50.589
Mean absolute error = 38.6744
Durbin-Watson statistic = 1.59176 (P=0.2323)
Lag 1 residual autocorrelation = 0.0975367

The StatAdvisor
---------------
   The output shows the results of fitting a
multiple linear regression model to describe the
relationship between Y and 3 independent
variables. The equation of the fitted model is

Y = 837.202 + 17.4761*X1 - 9.96119*X2 -
6.42127*X3

Since the P-value in the ANOVA table is less
than 0.01, there is a statistically significant
relationship between the variables at the 99%
confidence level.

The R-Squared statistic indicates that the model
as fitted explains 88.6278% of the variability
in Y. The adjusted R-squared statistic, which is
more suitable for comparing models with
different numbers of independent variables, is
86.4955%. The standard error of the estimate
shows the standard deviation of the residuals to
be 50.589. This value can be used to construct
prediction limits for new observations by
selecting the Reports option from the text menu.
The mean absolute error (MAE) of 38.6744 is the
average value of the residuals. The Durbin
Watson (DW) statistic tests the residuals to
determine if there is any significant
correlation based on the order in which they
occur in your data file. Since the P-value is
greater than 0.05, there is no indication of
serial autocorrelation in the residuals.

   In determining whether the model can be
simplified, notice that the highest P-value on
the independent variables is 0.2871, belonging
to X2. Since the P-value is greater or equal to
0.10, that term is not statistically significant
at the 90% or higher confidence level.
Consequently, you should consider removing X2
from the model.
```

Los resultados parciales muestran que un ajuste que considera a todas las variables, aunque los P-values muestran que X_2 es no significativa y que se puede obtener un mejor modelo al eliminarla del ajuste.

El análisis de varianza indica que al menos un coeficiente del modelo es diferente de cero, en el cuadro de parámetros se ve que realmente son 3 de 4.

El coeficiente de determinación indica una variación explicada del 88.62%

La estadística de Durban-Watson tiene un p-value de 0.2323 > 0.05, lo que indica de no se rechaza la hipótesis de independencia de los residuos (no autocorrelación).

En fin, la recomendación es leer con calma el texto del **StatAdvisor.**

```
Further ANOVA for Variables in the Order Fitted
-----------------------------------------------
Source    Sum of    Df   Mean      F-Ratio   P-Value
          Squares        Square
-----------------------------------------------
X1       258042.0    1  258042.0   100.83   0.0000
X2         1448.07   1    1448.07    0.57   0.4628
X3        59630.8    1   59630.8    23.30   0.0002
-----------------------------------------------
Model    319121.0    3

The StatAdvisor
---------------
   This table shows the statistical significance
of each variable as it was added to the model.
You can use this table to help determine how
much the model could be simplified, especially
if you are fitting a polynomial.
```

Esta última tabla muestra las significancias de cada variable explicativa al entrar al modelo, ya da idea de poder eliminar la variable X_2

95.0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	837.202	127.237	567.471	1106.93
X1	17.476	5.40612	6.01559	28.9366
X2	-9.96119	9.04575	-29.1374	9.21499
X3	-6.42127	1.33028	-9.24134	-3.60121

The StatAdvisor

This table shows 95.0% confidence intervals for the coefficients in the model. Confidence intervals show how precisely the coefficients can be estimated given the amount of available data and the noise which is present.

La tabla anterior da los intervalos de confianza para cada uno de los parámetros del modelo, esto es:

$$567.471 < \beta_0 < 1106.93$$

$$6.01559 < \beta_1 < 28.9366$$

$$-29.1374 < \beta_2 < 9.21499$$

$$-9.24134 < \beta_3 < -3.60121$$

Correlation matrix for coefficient estimates

	CONSTANT	X1	X2	X3
CONSTANT	1.0000	-0.6076	-0.2587	-0.9047
X1	-0.6076	1.0000	-0.5595	0.5749
X2	-0.2587	-0.5595	1.0000	0.0727
X3	-0.9047	0.5749	0.0727	1.0000

The StatAdvisor

This table shows estimated correlations between the coefficients in the fitted model. These correlations can be used to detect the presence of serious multicollinearity, i.e., correlation amongst the predictor variables. In this case, there are 2 correlations with absolute values greater than 0.5 (not including the constant term).

La tabla de correlaciones permite empezar a explorar posibles problemas de multicolinealidad. Se ve claramente que las correlaciones entre la constante y la variable X_3, y la variable X_1 y X_3 son mayores que 0.50 en valor absoluto, lo que indica colinealidad (no independencencia) entre las variables X_1 y X_3. Sin emabrgo, se debe correr todo el modelo nuevamente eliminando la variable X_2, antes de dar una conclusión definitiva.

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
1	506.0	599.025	-93.0253	-2.29

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Y deviates from a model fitted using all of the data except that observation. In this case, there is one Studentized residual greater than 2.0, but none greater than 3.0.

Tanto los residuales inusuales como los puntos de influencia permiten detectar observaciones o datos que influyen de manera significativa en el ajuste de un modelo lineal.

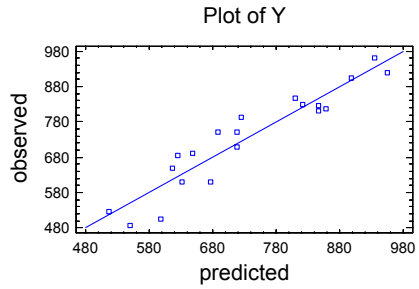
Influential Points

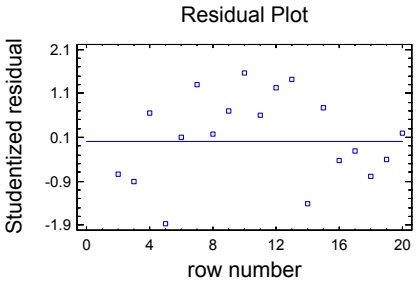
Row	Leverage	Mahalanobis Distance	DFITS
1	0.18594	3.16401	-1.09605
5	0.422505	12.2217	-1.60661
12	0.468013	14.888	1.14534

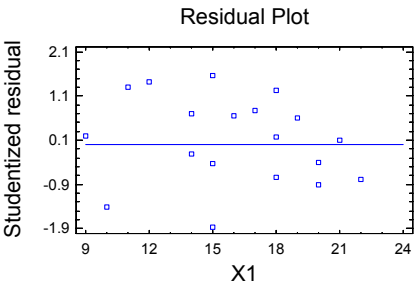
Average leverage of single data point = 0.2

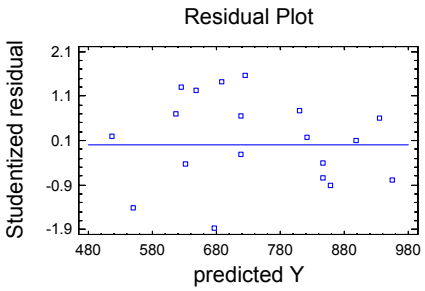
The StatAdvisor

The table of influential data points lists all observations which have leverage values greater than 3 times that of an average data point, or which have an unusually large value of DFITS. Leverage is a statistic which measures how influential each observation is in determining the coefficients of the estimated model. DFITS is a statistic which measures how much the estimated coefficients would change if each observation was removed from the data set. In this case, an average data point would have a leverage value equal to 0.2. There are no data points with more than 3 times the average leverage. There are 3 data points with unusually large values of DFITS.









Todos estos gráficos permiten analizar supuestos, en este caso no parece haber problemas con su cumplimiento.

INTERPRETACIÓN

El modelo a ajustar es del tipo

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Del ANOVA, se rechaza Ho: $\beta_i = 0$. En otras palabras, al menos una pendiente es significativa (diferente de cero), entonces si hay modelo.

De la tabla de coeficientes se tiene el siguiente modelo:

$$\hat{Y} = 837.202 + 17.476 x_1 - 9.961 x_2 - 6.421 x_3$$

Cuyo porcentaje de variación explicada es 86.5%. En esta misma tabla se tiene que el coeficiente de X_2 es no significativo (se puede considerar cero), de tal manera que se puede obtener un mejor modelo removiendo esta variable del modelo.

Con respecto a la multicolinealidad, este problema se presenta cuando entre las variables independientes existen relaciones lineales, es decir cuando las variables independientes dependen unas de otras (unas variables son combinaciones lineales de otras). Para su detección se utiliza, en primera instancia, la matriz de correlación de coeficientes, donde se considera que hay colinealidad si el valor absoluto de la correlación es mayor a 0.75. En este caso se tiene que el valor de correlación más alto es de 0.575, **no se tiene problemas de colinealidad**.

De este análisis surge la pregunta: ¿es mejor el modelo que no considera a X_2? Entonces se debe realizar el análisis de regresión mediante un **método de selección de variables.** Para esto:

- Dar un clic derecho sobre cualquier ventana de resultados.

- Del menú que aparece seleccionar Analysis Options.

- Elegir **Forward** (empieza con cero variables y se agregan las más significativas) o la opción **Backward** (empieza con todas las variables y elimina las menos significativas).

Al eligir Forward se tienen los siguientes resultados.

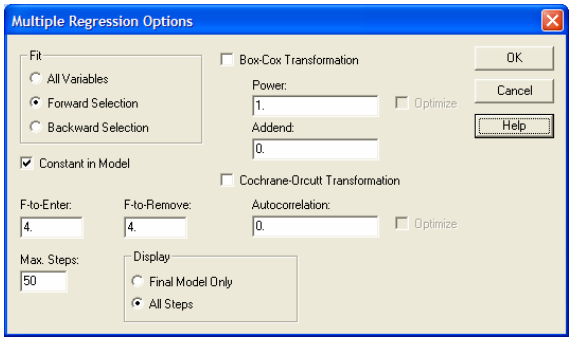


Figura133. Opciones para la selección de variables en la regresión múltiple.

```
Multiple Regression - Y

Multiple Regression Analysis
-----------------------------------------------
Dependent variable: Y
-----------------------------------------------
                     Standard   T
Parameter   Estimate   Error   Statistic   P-Value
-----------------------------------------------
CONSTANT    800.956  123.672   6.47645      0.0000
X1           14.1451 4.50862   3.13735      0.0060
X3          -6.31476 1.33503  -4.73006      0.0002
-----------------------------------------------

                Analysis of Variance
-----------------------------------------------
Source    Sum of    Df    Mean    F-Ratio    P-Value
          Squares         Square
-----------------------------------------------
Model    316018.0    2  158009.0  60.98      0.0000
Residual 44051.4   17    2591.26
-----------------------------------------------
Total    360069.0   19
(Corr.)

R-squared = 87.7659 percent
R-squared (adjusted for d.f.) = 86.3265 percent
Standard Error of Est. = 50.9044
Mean absolute error = 40.2901
Durbin-Watson statistic = 1.51985 (P=0.1702)
Lag 1 residual autocorrelation = 0.149211


Stepwise regression
-------------------
Method: forward selection
F-to-enter: 4.0
F-to-remove: 4.0


Step 0:
---------
0 variables in the model.  19 d.f. for error.
R-squared =   0.00%
Adjusted R-squared =   0.00%
MSE = 18951.0


Step 1:
     ---------
Adding variable X3 with F-to-enter = 75.1788
1 variables in the model.  18 d.f. for error.
R-squared =  80.68%
Adjusted R-squared =  79.61%
MSE = 3864.28


                    Step 2:
---------
Adding variable X1 with F-to-enter = 9.84298
2 variables in the model.  17 d.f. for error.
R-squared =  87.77%
Adjusted R-squared =  86.33%
MSE = 2591.26

Final model selected.


The StatAdvisor
---------------
   The output shows the results of fitting a
multiple linear regression model to describe the
relationship between Y and 3 independent
variables. The equation of the fitted model is

Y = 800.956 + 14.1451*X1 - 6.31476*X3

Since the P-value in the ANOVA table is less
than 0.01, there is a statistically significant
relationship between the variables at the 99%
confidence level.

     The R-Squared statistic indicates that the
model as fitted explains 87.7659% of the
variability in Y.    The adjusted R-squared
statistic, which is more suitable for comparing
models with different numbers of independent
variables, is 86.3265%.  The standard error of
the estimate shows the standard deviation of the
residuals to be 50.9044.  This value can be used
to construct prediction limits for new
observations by selecting the Reports option
from the text menu.  The mean absolute error
(MAE) of 40.2901 is the average value of the
residuals.    The Durbin-Watson (DW) statistic
tests the residuals to determine if there is any
significant correlation based on the order in
which they occur in your data file.  Since the
P-value is greater than 0.05, there is no
indication of serial autocorrelation in the
residuals.
   In determining whether the model can be
simplified, notice that the highest P-value on
the independent variables is 0.0060, belonging
to X1.  Since the P-value is less than 0.01, the
highest order term is statistically significant
at the 99% confidence level.  Consequently, you
probably don't want to remove any variables from
the model.


Further ANOVA for Variables in the Order Fitted
-----------------------------------------------
Source    Sum of    Df   Mean    F-Ratio    P-Value
          Squares        Square
-----------------------------------------------
X1        258042.0  1  258042.0  99.58     0.0000
X3         57975.4  1   57975.4  22.37     0.0002
-----------------------------------------------
Model     316018.0  2

The StatAdvisor
---------------
   This table shows the statistical significance
of each variable as it was added to the model.
You can use this table to help determine how
much the model could be simplified, especially
if you are fitting a polynomial.
```

95.0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	800.956	123.672	540.03	1061.88
X1	14.1451	4.5086	4.63275	23.6575
X3	-6.31476	1.33503	-9.13143	-3.49809

The StatAdvisor

This table shows 95.0% confidence intervals for the coefficients in the model. Confidence intervals show how precisely the coefficients can be estimated given the amount of available data and the noise which is present.

Correlation matrix for coefficient estimates

	CONSTANT	X1	X3
CONSTANT	1.0000	-0.9398	-0.9196
X1	-0.9398	1.0000	0.7447
X3	-0.9196	0.7447	1.0000

The StatAdvisor

This table shows estimated correlations between the coefficients in the fitted model. These correlations can be used to detect the presence of serious multicollinearity, i.e., correlation amongst the predictor variables. In this case, there is 1 correlation with absolute value greater than 0.5 (not including the constant term).

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
1	506.0	595.095	-89.0955	-2.12

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Y deviates from a model fitted using all of the data except that observation. In this case, there is one Studentized residual greater than 2.0, but none greater than 3.0.

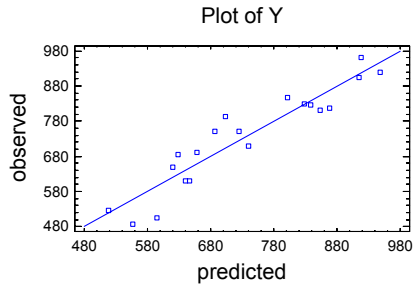
Influential Points

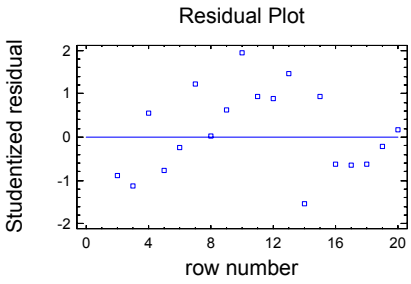
Row	Leverage	Mahalanobis Distance	DFITS
1	0.180963	3.02967	-0.998583
12	0.435016	12.912	0.780899

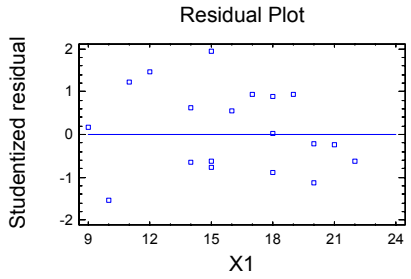
Average leverage of single data point = 0.15

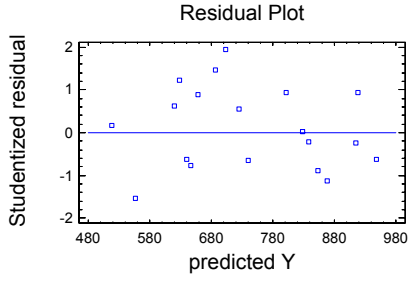
The StatAdvisor

The table of influential data points lists all observations which have leverage values greater than 3 times that of an average data point, or which have an unusually large value of DFITS. Leverage is a statistic which measures how influential each observation is in determining the coefficients of the estimated model. DFITS is a statistic which measures how much the estimated coefficients would change if each observation was removed from the data set. In this case, an average data point would have a leverage value equal to 0.15. There are no data points with more than 3 times the average leverage. There are 2 data points with unusually large values of DFITS.









INTERPRETACIÓN

En los resultados se puede notar como cambian los valores del cuadrado medio del error y de la R^2's en este último análisis. Pero lo más notorio es como cambia el comportamiento de los residuales al verificar los supuestos de normalidad.

Hay que definir cómo queda el nuevo modelo, su porcentaje de variación explicada y la significancia de los coeficientes.

EJERCICIOS

1. En un experimento realizado en el Laboratorio de Ciencia Básica I de la FES Zaragoza, se midió el enfriamiento del agua, a temperatura ambiente, con relación al tiempo, obteniéndose los siguientes datos:

Enfriamiento del H2O a Temperatura ambiente	
x (min)	y (ºC)
0.5	82
1	81
1.5	80
2	78
2.5	76.5
3	75.5
3.5	74
4	72.1
4.5	71
5	70.5
6	67.5
7	66
8	64
9	62.5
10	61
11	59
12	58
13	57
14	56.5
15	55

Obtenga la ecuación de regresión que mejor se ajusta a los datos y verifique supuestos.

2. En el mismo laboratorio se realizó otro experimento, donde se midió el enfriamiento del agua, en hielo, con relación al tiempo, obteniéndose los siguientes datos:

Efriamiento del H2O en el hielo	
x(min)	y (ºC)
0.5	81
1	78
1.5	74
2	69
2.5	68
3	49
3.5	44
4	40
4.5	34
5	31
6	28
7	21
8	17
9	15
10	14
11	13
12	12
13	10
14	9
15	9

Obtenga la ecuación de regresión que mejor se ajusta a los datos y verifique supuestos.

3. Para mejorar la calidad de un proceso de producción, es necesario comprender la capacidad del proceso Deming, 1982, Out of the Crisis). En un proceso particular de manufactura, la vida útil de una herramienta de corte está relacionada con la velocidad a la cual opera la herramienta. Es necesario entender esta relación para poder predecir cuando se debe reemplazar la herramienta y cuantas herramientas de repuesto se debe tener disponibles. Los datos en la siguiente tabla son datos reales obtenidos de una determinada marca de estas herramientas, usadas en el proceso de producción.
   a) Construya un diagrama de dispersión para los datos.
   b) Determine el modelo que mejor ajusta la vida útil vs. la velocidad de cortado.

Velocidad de cortado (metros/minu)	Vida útil (horas)
30	6.0
30	6.5
30	5.0
40	6.0
40	4.5
40	5.0
50	4.5
50	4.0
50	3.7

60	3.8
60	3.0
60	2.4
70	1.5
70	2.0
70	1.0

4. Con los siguientes datos,
   a) Identifique la ecuación de regresión.
   b) ¿Hay evidencia suficiente para establecer una relación lineal positiva entre x_1 con y?
   c) ¿Hay evidencia para establecer una relación lineal negativa entre x_2 con y?
   d) ¿Hay suficiente evidencia para establecer que el modelo de regresión es útil?

Y	X_1	X_2
100	7	28
104	11	27
106	13	29
109	15	31
115	16	26
118	18	24
123	20	20
131	23	18
136	25	22
139	28	20
150	33	19
151	34	17
153	39	14
158	41	12
159	42	14
164	44	13

5. Los siguientes datos son parte de un estudio grande conducido por el Dr. Rick Linthurst de la Universidad estatal de North Carolina como parte de la investigación de su tesis doctoral. El propósito de la investigación fue identificar las principales características del suelo que influyen en la producción de biomasa aérea en el pasto de pantano *Spartina alterniflora* en el estuario del Cabo de Miedo en North Carolina.

   Una fase de la investigación consistió en muestrear tres tipos de Spartina (áreas revegetativas "muertas" (DVEG), áreas de Spartina "enana" (SHRT) y áreas de Spartina "alta" (TALL)), en cada una de tres localidades (Oak Island (OI), Smith Island (SI) y Snows Marsh (SM)). Las muestras de sustrato de 5 lugares seleccionados al azar dentro de cada tipo de localidad-vegetación, dan un total de 45 muestras que fueron analizadas para 14 características del suelo para cada mes por varios meses. Además se midió la biomasa superficial para cada lugar de muestreo cada mes. Los datos usados en este estudio de caso involucraron solo los del mes de septiembre y las siguientes 5 medidas del sustrato.

   X_1 = Salinidad o/oo (SAL)
   X_2 = Acidez del agua (pH)
   X_3 = Potasio en ppm (K)
   X_4 = Sodio en ppm (Na)
   X_5 = Zinc en ppm (Zn)

   La variable dependiente Y es la biomasa aérea en $\frac{g}{m^2}$. Los datos para el mes de septiembre y las 6 variables se dan a continuación.

Obs.	Loc.	Tipo	BIO	SAL	pH	K	Na	Zn
1	OI	DVEG	676	33	5.00	1441.67	35184.5	16.4524
2	OI	DVEG	516	35	4.75	1299.19	28170.4	13.9852
3	OI	DVEG	1052	32	4.20	1154.27	26455.0	15.3276
4	OI	DVEG	868	30	4.40	1045.15	25072.9	17.3128
5	OI	DVEG	1008	33	5.55	521.62	31664.2	22.3312
6	OI	SHRT	436	33	5.05	1273.02	25491.7	12.2778
7	OI	SHRT	544	36	4.25	1346.35	20877.3	17.8225
8	OI	SHRT	680	30	4.45	1253.88	25621.3	14.3516
9	OI	SHRT	640	38	4.75	1242.65	27587.3	13.6826
10	OI	SHRT	492	30	4.60	1282.95	26511.7	11.7566
11	OI	TALL	984	30	4.10	553.69	7886.5	9.8820
12	OI	TALL	1400	37	3.45	494.74	14596.0	16.6752
13	OI	TALL	1276	33	3.45	526.97	9826.8	12.3730
14	OI	TALL	1736	36	4.10	571.14	11978.4	9.4058
15	OI	TALL	1004	30	3.50	408.64	10368.6	14.9302
16	SI	DVEG	396	30	3.25	646.65	17307.4	31.2865
17	SI	DVEG	352	27	3.35	514.03	12822.0	30.1652
18	SI	DVEG	328	29	3.20	350.73	8582.6	28.5901
19	SI	DVEG	392	34	3.35	496.29	12369.5	19.8795
20	SI	DVEG	236	36	3.30	580.92	14731.9	18.5056
21	SI	SHRT	392	30	3.25	535.82	15060.6	22.1344
22	SI	SHRT	268	28	3.25	490.34	11056.3	28.6101
23	SI	SHRT	252	31	3.20	552.39	8118.9	23.1908
24	SI	SHRT	236	31	3.20	661.32	13009.5	24.6917
25	SI	SHRT	340	35	3.35	372.15	15003.7	22.6758
26	SI	TALL	2436	29	7.10	525.65	10225.0	0.3729
27	SI	TALL	2216	35	7.35	563.13	8024.2	0.2703
28	SI	TALL	2096	35	7.45	497.96	10393.0	0.3705
29	SI	TALL	1660	30	7.45	458.38	8711.6	0.2648
30	SI	TALL	2272	30	7.40	498.25	10239.6	0.2105
31	SM	DVEG	824	26	4.85	936.26	20436.0	18.9875
32	SM	DVEG	1196	29	4.60	894.79	12519.9	20.9687
33	SM	DVEG	1960	25	5.20	941.36	18979.0	23.9841
34	SM	DVEG	2080	26	4.75	1038.79	22986.1	19.9727
35	SM	DVEG	1764	26	5.20	898.05	11704.5	21.3864
36	SM	SHRT	412	25	4.55	989.87	17721.0	23.7063
37	SM	SHRT	416	26	3.95	951.28	16485.2	30.5589
38	SM	SHRT	504	26	3.70	939.83	17101.3	26.8415
39	SM	SHRT	492	27	3.75	925.42	17849.0	27.7292
40	SM	SHRT	636	27	4.15	954.11	16949.6	21.5699
41	SM	TALL	1756	24	5.60	720.72	11344.6	19.6534
42	SM	TALL	1232	27	5.35	782.09	14752.4	20.3295
43	SM	TALL	1400	26	5.50	773.30	13649.8	19.5880
44	SM	TALL	1620	28	5.50	829.26	14533.0	20.1328
45	SM	TALL	1560	28	5.40	856.96	16892.2	19.2420

a) Ajuste un modelo de regresión lineal múltiple para los datos.
b) Analice la adecuación del modelo mediante un ANOVA.
c) Realice el diagnóstico del modelo (normalidad y homoscedasticidad mediante residuos, casos extraordinarios (outliers) y medidas de influencia, multicolinealidad y autocorrelación) e identifique las inadecuaciones que presente.
d) Analice los datos mediante la selección de variables y proponga el modelo más adecuado.

NOTA FINAL: Un ejercicio interesante consiste en realizar un análisis de varianza y a partir de los resultados hacer un análisis de regresión. De manera que no sólo se detecten las variables significativas sino que también se obtenga un modelo que describa el comportamiento de los datos.

BIBLIOGRAFÍA

Cervantes, S. A., Rivera, G. P. y De la Paz L. J. M., (2004), *SPSS. Una Herramienta para el Análisis Estadístico de Datos*, FES Zaragoza UNAM, México.

Charterjee, S. y Price, B. (1991), *Regression Analysis by Example*, 2a. edición, John Wiley and Sons, Inc., N. Y., U.S.A

Devore, J. L. (2001), Probabilidad y Estadística para Ingeniería y Ciencias. 5ª. edición. Ed. Thomson Learning, México.

Draper, N. R. y Smith, H. (1981), *Applied Regression Analysis*, 2a. ed., John Wiley and Sons, Inc., N.Y., U.S.A.

Freund, J. E. y Simon, G. A. (1992), *Estadística Elemental*, Prentice Hall, Inc. U.S.A.

Marques, M. J. (2004)), *Probabilidad y Estadística para Ciencias Químico Biológicas*, 2ª. edición, FES Zaragoza UNAM, México.

McClave, J. T., Dietrich, F. H. II and Sincich, T., (1997), *Statistics*, 7th. edition. Prentice Hall, Inc., New Jersey, U.S.A.

Montgomery, D. C. y Peck, E. A. (1992), *Introduction to Linear Regression Analysis*, 2a. edición, John Wiley and Sons, Inc. N.Y., U.S.A.

Pérez Cesar, (1996), *Econometría y Analisis Estadístico Multivariable con Statgraphics*. Técnicas Avanzadas, RA-MA editorial, Madrid, España.

Rawlings, J. O. (1988), *Applied Regression Analysis, A research Tool*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California, U.S.A.

Seber, G. A.F. (1977), *Linear Regression Analysis*. John Wiley and Sons, Inc. N.Y., U.S.A.

Statgraphics Plus 5.0, (1994-200), *Ayuda y tutorial integrado al software*.

Velleman, P. F. Y Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis.* Duxbury Press, Boston, Massachusets, U.S.A.

Análisis Estadístico.
Un enfoque práctico con Statgraphics.

1era. edición
Se imprimió en el Laboratorio de Aplicaciones
Computacionales de la FES Zaragoza.
Con un tiraje inicial de 100 ejemplares
y una versión electrónica en CD-ROM

Universidad Nacional Autónoma de México
Facultad de Estudios Superiores Zaragoza