Emilio Letón Molina
Alejandro P. Marino

G-Stat 2.0

Programa de Análisis Estadísticos

Manual del Programa, Documentación y Ejemplos

Versión del manual 2.0

Base de Datos
Estadística Descriptiva y de Estimación
Pruebas Estadísticas
Epidemiología
Diagnóstico
Técnicas Multivariantes
Gráficos

www.g-stat.es

www.e-biometria.com

Autores

Emilio Letón Molina
Alejandro P. Marino

Dpto. Biometría, GlaxoSmithKline S.A., Tres Cantos, Madrid

Edita GlaxoSmithKline S.A.

ISBN:84-607-5171-6. Registro Legal: M-37418-2002

Consultores informáticos y estadísticos

Sacha Arozarena, Alfonso Palacios, Álvaro Trigo, Gema Núñez (ASI, Madrid); Llorenç Badiella (Universitat Autònoma, Barcelona)

Condiciones de utilización

La distribución de este programa se realiza sobre la base del concepto de distribución gratuita. Los propietarios renuncian a los derechos de comercialización aunque mantienen los derechos de propiedad intelectual e industrial. La cesión libre y gratuita de *G-Stat* no incluye la cesión de los derechos de propiedad del programa. Por ello los usurarios no podrán:

- Modificar los programas contenidos en *G-Stat* ni realizar versión alguna del código fuente por decompilación u otro método.
- Alterar, modificar o adaptar la documentación, el programa o el aspecto de las pantallas.

Los propietarios no se hacen responsables de los daños y perjuicios, directos o indirectos, especiales o incidentales, que se deriven del uso y utilización, debida o indebida, del programa o de la documentación que se adjunta. No se permite la reproducción total o parcial de esta publicación, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio, ya sea electrónico, mecánico, por fotocopia, por registro u otros métodos, ni su préstamo, alquiler o cualquier otra forma de cesión de uso de este Manual, sin el permiso previo y por escrito de los propietarios del programa.

Actualizaciones de este manual

En www.g-stat.es se encuentrán las versiones actualizadas de este manual.

Contenido

Generalidades

Descripción

El departamento de Biometría de GSK ha desarrollado el programa G-Stat 2.0, un programa completo de análisis estadísticos.

G-Stat es un programa estadístico que se ha desarrollado en Java y que se puede instalar bajo Windows, Unix, Linux y Macintosh. Está diseñado para que el usuario pueda realizar, de una forma rápida y fácil, análisis estadísticos con resultados tanto gráficos como numéricos. Para el manejo del programa no es necesario tener amplios conocimientos estadísticos. El programa se maneja por menús y no requiere programación. Incluye base de datos, gráficos, estadística descriptiva, técnicas de estimación y pruebas estadísticas univariantes y bivariantes, paramétricas y no paramétricas, técnicas de diagnóstico y de epidemiología. Asimismo, incluye técnicas multivariantes esenciales como el análisis de la varianza, regresión lineal múltiple, regresión logística y regresión de Cox.

Requerimientos del equipo

Los requisitos recomendados para que G-Stat funcione de forma satisfactoria son:

- Procesador Pentium 350 MHz o similar
- 128 MB de memoria RAM
- 100 MB de espacio libre en disco duro
- Lector de CD
- Pantalla VGA (1024 x 768) con 256 colores de resolución
- Impresora configurada

Instalación

Ejecutar el archivo Install.htm y seguir las instrucciones. Información actualizada sobre G-Stat se encuentra en:

www.g-stat.es

www.e-biometria.com

Nota

G-Stat es un programa multisistema operativo. Muchos usuarios encontrarán características similares a los programas en Windows, pero al ser una aplicación programada en Java su apariencia y uso puede diferir.

Mejoras con respecto a las versiones 1.x

Técnicas nuevas incorporadas

- Gráficos de bloques de medias y desviaciones típicas.
- Series temporales: gráficos, medias móviles, alisado exponencial, alisado exponencial con doble parámetro de Holt-Winters.
- Estadísticos de asociación para datos agrupados.
- Prueba z-proporción para datos sin agrupar
- Prueba de Shapiro-Wilk.
- Levene.
- Chi-Cuadrado para datos agrupados.
- Chi-Cuadrado de tendencia lineal.
- Chi-Cuadrado de tendencia lineal para datos agrupados.
- Prueba de Fisher para datos agrupados.
- Prueba de McNemar para datos agrupados.
- Epidemiología para datos agrupados.
- Mantel-Haenszel para datos agrupados.
- Coeficiente Kappa de concordancia.
- Breslow-Day.
- Técnicas de Diagnóstico.
- Diagnóstico para datos agrupados.
- Curvas ROC.
- Prueba de Log-Rank.
- Coeficiente de correlación intraclase.
- Anova Un Factor para datos agrupados.
- Comparaciones múltiples a posteriori no paramétricas de Dunn.
- Regresión Logística hacia adelante y hacia detrás.
- Cox-Snell y Nagelkerke.
- Regresión de Cox hacia adelante y hacia detrás.

Instalación

- Incorporación de un instalador o asistente que permite la instalación más fácil en los principales sistemas operativos: Win, Mac, Linux y Unix.

- Desaparece la pantalla negra cuando se instala en Windows.

General

- Se ha incorporado un reloj para indicar que una técnica se está procesando.
- En los menús se incluyen pistas abreviadas de utilización de las técnicas.
- Se ha incorporado un botón de Imprimir en todas las pantallas de resultados.
- Se incluyen instrucciones para importar datos de Excel y otros programas.
- Control sobre el número de decimales de cada variable.
- Se han ampliado las ayudas “on-line” del programa en todos los botones de Ayuda.
- Avisos en operaciones no adecuadas.
- Aumento de la robustez: controles de funcionamiento para casos y ficheros extremos.

Regresión Lineal Múltiple

- Se ha mejorado la velocidad del cálculo de los residuos “jacknife”.

Regresión Logística

- Rediseño del código de programación incorporando optimización en el cálculo matricial del producto por matrices diagonales para incrementar la rapidez de su ejecución.
- Estandarización y desestandarización interna de las variables para eliminar errores de redondeo.
- Separación del cálculo de la verosimilitud para evitar realizar operaciones no factibles con logaritmos neperianos.
- Se detectan situaciones anómalas de convergencia: separación y cuasiseparación.
- Se ha mejorado el código para asegurar la convergencia mediante el método de “half-step”.
- Nuevas opciones para realizar el modelo: con constante y sin constante, selección en bloque , hacia delante y hacia detrás, punto de corte, p-para-entrar, p-para-salir, número máximo de iteraciones, mostrar las iteraciones, punto de corte.
- Se incluyen las predicciones del modelo.
- Nuevos estadísticos para el término constante.

Regresión de Cox

- Estandarización y desestandarización interna de las variables para eliminar errores de redondeo.
- Separación del cálculo de la verosimilitud para evitar realizar operaciones no factibles con logaritmos neperianos.
- Se ha mejorado el código para asegurar la convergencia mediante el método de “half-step”.
- Nuevas opciones para realizar el modelo: selección en bloque, hacia delante y hacia detrás, punto de corte, p-para-entrar, p-para-salir, número máximo de iteraciones, mostrar las iteraciones.

Corrección de errores

- En el manejo de la base de datos: insertar y eliminar registros.
- Etiquetas e impresión de los gráficos.
- Homogeneización entre clases e histogramas.
- Distribución Normal para valores negativos.
- Contraste de hipótesis de una proporción.
- Rangos signados, Mann-Whitney y Wilcoxon Pareados en situaciones extremas.
- Fisher se ha corregido en el caso de que se utilice en su cálculo factoriales de números elevados.
- Se han corregido algunos errores en el cálculo de los estadísticos D+ y D- de Kolmogorov, aunque no afectaban al cálculo del p-valor Lilliefors corregido.
- Etiqueta de los grados de libertad en el Anova Un Factor con Bloques
- Regresión Lineal Múltiple sin constante hacia delante y hacia detrás.

Dominio Web Propio

G-Stat cuenta con dominio propio: www.g-stat.es donde se encuentra información actualizada del programa.

Notas sobre el manual

Este manual no es un libro de estadística. No se pretende que los usuarios aprendan estadística con la información aquí contenida. La principal función consiste en proporcionar la formulación empleada en la programación.

El Manejo del programa es en su mayor parte autoexplicativo. En muchas ocasiones, al igual que en el programa se repiten técnicas, en este manual la información es redundante, pero facilita el manejo.

Mucha de la ayuda contenida en esta manual está incluida en los botones de ayuda de las pantallas de resultados.

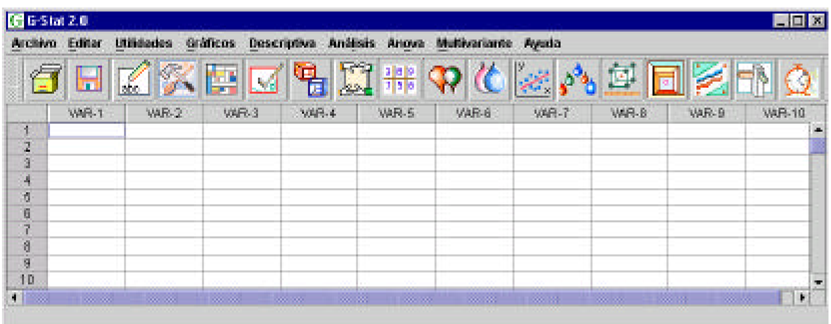
No se incluyen ejemplos en todas las técnicas, sólo en las que se ha considerado de más interés o más didácticas.

Como norma general conviene usar siempre ficheros con al menos dos líneas de datos y que no haya variables constantes. Las opciones de Multivariante en algunas ocasiones pueden tardar cierto tiempo si hay muchos datos.

Arranque de G-Stat

El programa G-Stat es compatible con otros programas y está validado en modo multitarea. Sin embargo, los salvapantallas y los antivirus (u otros programas residentes) pueden provocar, ocasionalmente, errores de protección general según la plataforma, el equipo y el software con el que se esté trabajando.

Al arrancar el programa aparece la pantalla principal, compuesta por ***Barra de título***, ***Menú Principal*** y ***Barra de Iconos***. Ésta es la pantalla para la introducción y manejo de datos.



Pantalla principal del programa.

Barra de título: Está situada en la parte superior de la ventana y contiene el nombre del programa.

Menú principal: Está debajo de la *barra de título*. Éste llevará a los submenús con las opciones principales del programa.

***Barra de iconos**:* Se encuentra después del *menú principal y* dispone de una serie de iconos que permiten realizar ciertas operaciones directamente, sin necesidad de utilizar los menús. Al posicionar el puntero encima del icono se identifica la opción.

Los iconos activan las siguientes opciones:

- Abrir Archivo
- Guardar Archivo
- Editor de Texto
- Transformar
- Ejemplo
- Validación
- Contraste de hipótesis de dos proporciones
- McNemar
- Chi-Cuadrado
- t-Student
- Man-Whitney (Wilcoxon)
- Regresión Lineal Simple
- Kruskal-Wallis
- Anacova
- Anova Dos Factores
- Regresión Múltiple
- Regresión Logística
- Regresión de Cox
- Manual de G-Stat *
- Salir *

*Visible en pantallas superiores a 17 pulgadas o resoluciones superiores a 800 x 600 pixeles.

Menú Principal

Navegación

El menú principal está compuesto por submenús que contienen las funciones más usuales de los programas utilizados en Windows y aquellas que proporciona el programa G-Stat. El acceso a las opciones de los menús se realiza por puntero o mediante combinaciones de teclas. Todos los menús contienen submenús.

La navegación por el programa permite abrir simultáneamente varias ventanas. Sin embargo, cualquier cambio en la ventana de datos no tiene efecto en las ventanas de resultados abiertas. Para actualizar los resultados hay que actualizar sus ventanas.

En muchas de las pantallas de resultados se encuentran botones que abren las ventanas de opciones de análisis. El programa no realiza las operaciones si las opciones no se confirman mediante el botón Aceptar.

Las ventanas de resultados contienen pestañas con diferentes subanálisis. Generalmente las opciones de las pestañas son independientes, de tal manera que el usuario debe confirmar en cada pestaña si las opciones son las adecuadas y no confiarse en las opciones por defecto del programa ni en cambios previos.

Botones

En la mayoría de las ventanas del programa G-Stat se encuentran los siguientes botones:

Aceptar: Cierra la ventana aceptando lo realizado.

Cancelar: Cierra la ventana sin aceptar lo realizado.

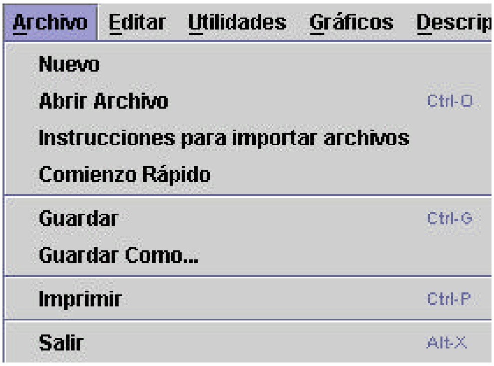
Guardar: Abre el cuadro de diálogo para guardar el gráfico en un fichero.

Imprimir: Muestra el cuadro de diálogo de impresión. Imprime el gráfico.

Opciones: Opciones relacionadas con la técnica estadística o gráfico.

Ayuda: Da información complementaria para interpretación de resultados.

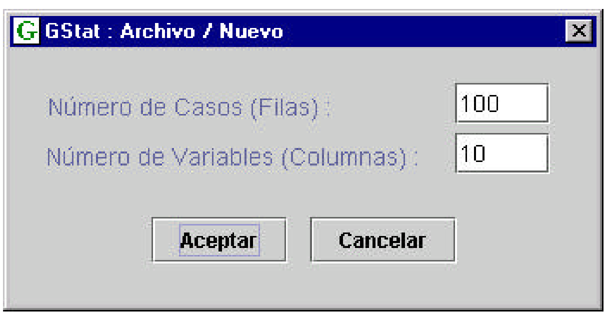
Menú Archivo



Activar la opción **Archivo** del menú principal o mediante Alt+A. Este menú contiene las opciones para el manejo de la base de datos.

Nuevo

Crea una nueva base de datos vacía. Aparece una pantalla que pregunta *si se desea guardar el archivo actual*. Después de aceptar, se abrirá una ventana donde se tiene que introducir el número de casos (Filas) y el número de variables (Columnas). Por defecto G-Stat siempre crea una base de datos con 100 filas y 10 columnas. Introducido el número de filas y columnas pulsar el botón *Aceptar*.



Ventana de definición de la estructura de la base de datos nueva.

Entrada de datos

Para introducir los datos de cada registro, hacer clic en la casilla correspondiente. Para moverse por la base de datos ir directamente con el puntero o con los cursores del teclado. Para editar cualquier valor hacer doble clic en la casilla correspondiente.

- Disposición de los datos

La hoja de datos está estructurada en una cuadrícula de filas y columnas: las filas corresponden a los individuos o casos y las columnas a las variables.

- Tipo de variables

El programa admite variables numéricas y alfanuméricas. Sin embargo, para la realización de la mayoría de las pruebas y técnicas estadísticas se exige que las variables sean numéricas. Es recomendable, por tanto, la consignación de las variables como numéricas aunque su naturaleza sea nominal o dicotómica. Es posible recodificar valores de variables nominales o dicotómicas mediante la opción "Recodificar" del menú "Utilidades". El programa trata las variables fecha como variables nominales.

- Decimales

Se utiliza el punto como separador decimal. Se puede emplear tanta precisión como se desee para cada variable. El número de decimales no tiene por qué ser el mismo para los diferentes valores numéricos de una variable. Si por error se emplea la coma o se importa un fichero de datos que emplea la coma como separador decimal, los datos serán tratados como alfanuméricos.

- Valores "Missing"

El programa admite valores “missing” o “valores faltantes”. Un valor missing en una variable no anula el registro, excepto en las técnicas estadísticas que operan simultáneamente con varias variables.

- Menú contextual

El botón derecho activa el menú contextual en la hoja de datos activando las opciones:

- Cambiar Nombre
- Número de Decimales
- Insertar Variable
- Eliminar Variable
- Orden Ascendente
- Orden Descendente
- Recodificar

La mayoría de estas opciones se encuentran también disponibles en los menús “Editar” y “Utilidades”.

- Formato

Es posible variar la anchura de las columnas situando y moviendo el cursor entre las líneas de separación de las variables. El tipo de variable y el número de decimales es reconocido automáticamente por el programa.

- Editar un dato

Hacer doble clic sobre la celda. El nuevo dato sobreescribirá el antiguo. Para editar parcialmente un valor hacer doble clic y, tras una pausa, hacer un clic sobre la celda. Para añadir, hacer un solo clic sobre la celda.

- Navegación

Utilizar las teclas de posición para moverse por la cuadrícula. Alternativamente utilizar el cursor del ratón y las teclas de avance rápido de página. Mediante las teclas “Ctrl-Fin” se posiciona en el final de la base de datos. Mediante las teclas “Ctrl-Inicio” se posiciona en el comienzo de la base de datos. La tecla "Enter" mueve el cursor a la celda inferior. El tamaño de la pantalla de la hoja de datos es ajustable mediante los cursores activos en los laterales y en el extremo inferior derecho.

Abrir Archivo

(Ctrl+O). Esta opción permite abrir ficheros mediante el cuadro de diálogo de apertura de ficheros. Al seleccionar esta opción aparecerá una pantalla que pregunta *si se desea guardar el archivo actual*.

En el cuadro de diálogo *Abrir archivo*, se selecciona la unidad, el directorio y el nombre de fichero. Hecha la selección del archivo, pulsar el botón *Aceptar* e introducir las características del fichero.

Al abrir una base de datos, automáticamente, se cerrará el fichero que estaba activo hasta ese momento.

El programa contiene una base de datos "*ejemplo*" que se copia automáticamente al realizar la instalación y que se abre mediante el menú Utilidades o mediante el icono correspondiente.

Instrucciones para importar archivos

Para importar datos desde Excel, realizar los siguientes pasos:

1.- Desde Excel, hacer Archivo / Guardar como: guardar como tipo: texto (delimitado por tabulaciones). Hay que tener en cuenta que el separador decimal debe ser el punto en lugar de la coma. Esto se cambia en Menú de Inicio / Configuración / Panel de Control / Configuración Regional / Pestaña: Número / Campo: Signo Decimal.

2.- Abrir el archivo de texto con G-Stat mediante Archivo / Abrir Archivo.

Para importar datos desde Access, SPSS o cualquier otro programa, realizar los siguientes pasos:

1.- Exportar los datos a texto y seguir el paso 2 anterior o exportar los datos a Excel y seguir los pasos 1 y 2 anteriores.

Comienzo Rápido

Se incluye, a modo de ejemplo, los pasos que habría que seguir para realizar la prueba de comparación de medias t-Student. El objetivo de este menú es facilitar una guía rápida (en un minuto) de las posibilidades del programa.

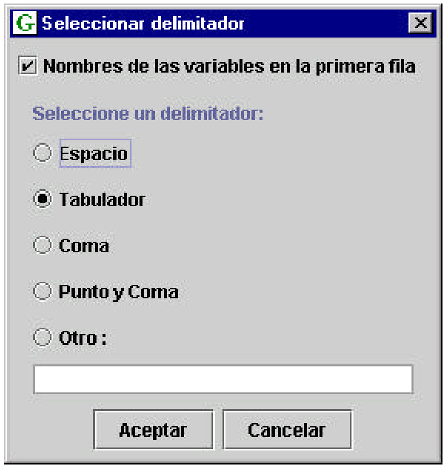
Guardar

(Ctrl+G). Guarda las últimas modificaciones de la base de datos que está en uso. Aparece el cuadro de diálogo de características del fichero. Seleccionadas las características hacer clic en el botón *Aceptar*. Si la base de datos es nueva, no estará guardada aún y aparecerá la ventana de *Guardar como*.

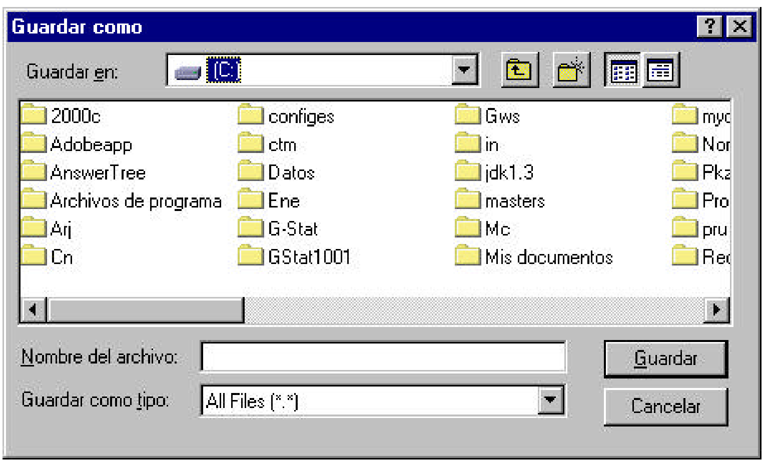
Guardar Como ...

Guarda el fichero en otra ubicación. Una vez elegidas las características del fichero, aparece la ventana *Guardar como* donde se selecciona la unidad y el directorio donde se desea guardar la base de datos y su nombre.

Guardar la base de datos en otra ubicación, por ejemplo, en Mis Documentos.



Opciones de "Guardar Como" con las características de los registros.



Ventana "Guardar como" del programa.

Imprimir

(Ctrl+P). Imprime la base de datos activa. Se abre la ventana de impresión del sistema operativo (S.O.). Se selecciona la impresora y el número de copias a realizar.

Salir

(Alt+X). Permite salir del programa. Al salir del programa se muestra un mensaje por si se desea guardar las últimas modificaciones realizadas. Si se selecciona el botón *Sí,* saldrá la ventana de características de la base de datos, una vez realizada la selección se acepta y se sale de la aplicación. Con el botón *No,* se sale directamente del programa sin guardar los cambios.

Menú Editar

Activar la opción **Edición** del menú principal o mediante Alt+E. Este menú contiene las opciones relativas a la edición de texto.

G-Stat permite la edición y grabación de datos pero no es un programa especialmente diseñado para la gestión avanzada de bases de datos. Se recomienda realizar las operaciones previas en la base de datos con otras herramientas más potentes, y una vez validada la base de datos, exportarla a G-Stat mediante su conversión en fichero texto separado por tabuladores (opción recomendada) y con el nombre de las variables en primera fila.

El programa añade automáticamente filas vacías en la base de datos cuando se añaden o suprimen líneas (registros). Esto no altera el funcionamiento del programa ya que sólo considera los registros que contienen información en alguna de las variables.

Cortar

(Ctrl+X). Permite cortar uno o varios registros para eliminarlos de la base de datos activa, para insertarlos en otras celdas de la base de datos o para pegarlos en otra base de datos. Se inserta o pega a través de la opción Pegar. Seleccionar la o las casillas a cortar. Activar la opción Cortar. Se borrarán los registros seleccionados y se copiarán en el portapapeles.

Copiar

(Ctrl+C). Permite realizar una copia de uno o varios registros para pegarlo en la misma base de datos o en otra. Se seleccionan los registros y se activa la opción Copiar. La información se copiará en el portapapeles.

Pegar

(Ctl+V). Pega la información guardada en el portapapeles. Se sitúa el cursor en la casilla donde se va a pegar el registro. En el caso de que sea más de uno se selecciona la primera casilla.

Eliminar Variable / Registro

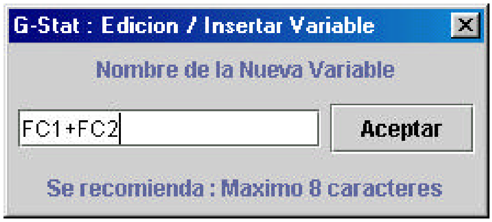
Elimina uno o varios registros o variables de la base de datos. Los pasos a seguir son los mismos en estas dos últimas opciones, se selecciona con el puntero la(s) variable(s) o registro(s) a eliminar. Se activa la opción correspondiente y aparece un cuadro de diálogo para asegurar que se desea eliminar la variable o el registro. Se pulsa *Aceptar* para eliminar.

Insertar Variable

Inserta una variable nueva a la base de datos activa. Se sitúa el cursor en la variable anterior a la que se va insertar. Se selecciona la opción Insertar. Aparece una ventana en la cual se tiene que introducir el nombre de la nueva variable. Se crea la nueva variable vacía.

Insertar la variable FC1+FC2 después de la variable FC2.

Se posiciona el cursor en la variable FC2FC1. Se selecciona la opción Insertar Variable.



Ventana Insertar Variable.

Esto permite nominar una nueva variable, pero los datos se deberán grabar, importar o generar por la opción transformación.

Insertar Registro

Inserta un nuevo registro en la base de datos abierta. Se sitúa el cursor en el registro siguiente al que se desea insertar. Se selecciona la opción Insertar Registro. Se crea el nuevo registro en blanco. Esta opción sirve para insertar más de un registro a la vez, señalando varias filas en el marcador a la izquierda.

Ordenar

Se coloca el puntero en la variable por la cual se quiere ordenar la base de datos. Ordena de forma ascendente o descendente la base de datos activa en relación a la variable seleccionada.

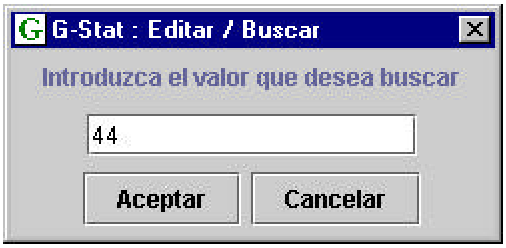
Compactar Tabla

Elimina los registros vacíos intermedios de la base de datos activa. No elimina los registros finales vacíos permanentemente presentes. Estos registros finales vacíos no son nunca considerados en los cálculos estadísticos.

Buscar

Busca un valor en una variable de la base de datos abierta. Colocar el cursor en la variable donde se va a buscar el valor. Seleccionada la opción Buscar, aparece un cuadro de diálogo donde se introduce el valor a buscar. Pulsar el botón *Aceptar* y el cursor se situará en el valor encontrado, si no existe dicho valor, aparecerá un cuadro de diálogo que indica que no se ha encontrado el valor.

Identificar el registro del paciente que tiene, en la variable FC2FC1, el valor 44.



Ventana Buscar.

Ir a Registro

Busca un registro en la base de datos actual. Aparece una ventana donde se introduce el número de orden del registro a buscar. Al pulsar el botón *Aceptar* se seleccionará el registro buscado. Si el registro no existe aparece un cuadro de diálogo que indica que el registro introducido no es válido.

Menú Utilidades

Activar la opción **Utilidades** del menú principal o mediante Alt+U. Este menú contiene las opciones: abrir un editor de textos independiente, modificar los datos mediante transformaciones y recodificaciones, realización de filtros de la base de datos y un ejemplo.

Editor de texto

Abre el editor de texto del programa. Al seleccionar esta opción aparece el editor de texto del programa con un único menú, **Archivo**. Dicho menú contiene las funciones básicas para el tratamiento de ficheros: *Nuevo, Abrir, Insertar, Guardar, Guardar como, Imprimir y Salir*. Para imprimir los resultados del programa, se puede copiarlos previamente al editor o usar directamente el botón Imprimir en los resultados. Esta utilidad es similar al editor de texto del sistema operativo.

Transformar

Permite generar variables mediante la transformación de variables numéricas ya creadas. En la ventana *Transformar*, aparece un recuadro con todas las variables cuantitativas de la base de datos. La o las variables que se utilizan para la transformación se llevarán a los recuadros blancos. La transformación puede estar compuesta por una variable y un número. Se elige el *operador* a utilizar del menú de los operadores y se identifica la *variable.*

Las transformaciones sobre datos faltantes ("missing") dan un resultado faltante, incluso aunque las celdas destino estuviesen previamente rellenas.

Anteriormente se ha creado la variable FC1+FC2, ahora se desea rellenarla con la suma de las variables FC1 y FC2.

El resultado será la variable FC1+FC2.

Cuadro de diálogo Transformar.

Recodificar

Permite cambiar automáticamente uno o varios valores, tanto numéricos como alfanuméricos, de una variable. Colocar el cursor en la variable a recodificar, que puede ser tanto cuantitativa como cualitativa. Aparece la ventana de recodificar compuesta por dos columnas: *Antes*, con los diferentes valores de la variable a recodificar y *Después*, donde inicialmente aparecen los mismos valores. En la columna *Después* se pueden ir definiendo las modificaciones de los valores o eliminarlos si se desea. Para que el programa considere todas las modificaciones hay que mover el cursor a otra celda después de la última entrada. Definidas las modificaciones, pulsar el botón *Aceptar*.

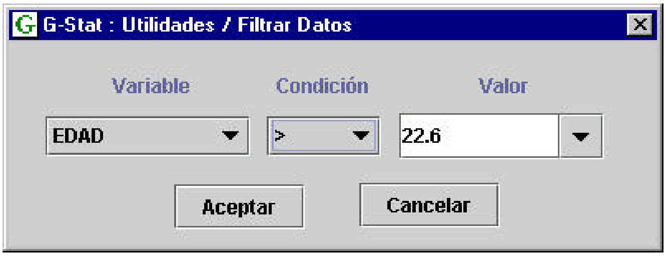
Se desea codificar la variable Sexo en 0 y 1, pero en la base de datos aparece esta variable codificada en 1 y 2. Se recodificará la variable de forma que donde antes había un 1 se introduce un 0 y donde había un 2 se introduce un 1.

Ventana de la opción recodificar.

Filtrar Datos

Permite seleccionar individuos a partir de un rango específico correspondiente a una variable. Seleccionada la opción se muestra la ventana para filtrar datos. Se selecciona la variable, la condición lógica utilizada para filtrar los datos y el valor del filtro. Pulsar *Aceptar* para finalizar.

Se desea realizar un subestudio sólo con las personas mayores de 22.6 años. Se filtrarán los datos según esta condición.



Ventana de filtrar datos mediante la variable Edad.

Ejemplo

Abre una base de datos que contiene variables de diferentes tipos y permite la realización de casi todas las técnicas y análisis estadísticos del programa. Seleccionada la opción *Ejemplo* aparece el cuadro de diálogo de guardar, aunque aún no se haya abierto ninguna base de datos.

El ejemplo incorporado en el programa se utiliza en este manual. El fichero es *pulsofar6.gst*, y se encuentra en la carpeta de instalación del programa. El fichero contiene 40 registros correspondientes a otros tantos sujetos. Las variables incluidas son:

IB: Número de identificación

Sexo: 1=Hombre; 2=Mujer

Fumador: 1=sí; 2=no

Edad: Edad en años

FC1: Frecuencia cardiaca antes del ejercicio

FC2: Frecuencia cardiaca después del ejercicio

FC2FC1: Incremento de la frecuencia cardiaca

Status: Nivel de entrenamiento físico 1, 2 ó 3

Farmaco: 1=Fármaco1; 2=Fármaco2

Los datos son ficticios pero plausibles.

Menú Gráficos

Activar la opción **Gráficos** del menú principal o mediante Alt+G. Este menú contiene las opciones necesarias para la realización de gráficos. Las opciones de este menú están separadas en dos grupos: gráficos univariantes y gráficos bivariantes. Los códigos (a) o (b) indican que este tipo de gráficos son adecuados para variables cualitativas y los códigos (y) o (x) para variables cuantitativas.

La separación por barras verticales indica un modelo que asume que las variables a la izquierda de la barra representan las variables explicativas o independientes y a la derecha la variable respuesta o dependiente.

Las pantallas de gráficos tienen cuatro botones: *Guardar, Imprimir, Opciones* y *Ayuda.*

Barras (a)

Crea un gráfico de barras para una variable cualitativa o discreta. Los gráficos de barras se construyen de forma que la longitud / altura de la barra corresponde a la frecuencia absoluta para cada uno de los niveles de la variable. El orden y el color de las barras dependen de la disposición.

Manejo del programa

Identificar la variable a analizar y activar la pestaña **Barras** donde aparece el gráfico de barras en una nueva ventana.

Opciones:

- La cabecera, orientación del gráfico, el título del eje X, su escala (absoluta/frecuencias, relativa/porcentajes), mínimo, máximo e incremento.

Obtener la distribución por sexos de la base de datos del ejemplo.

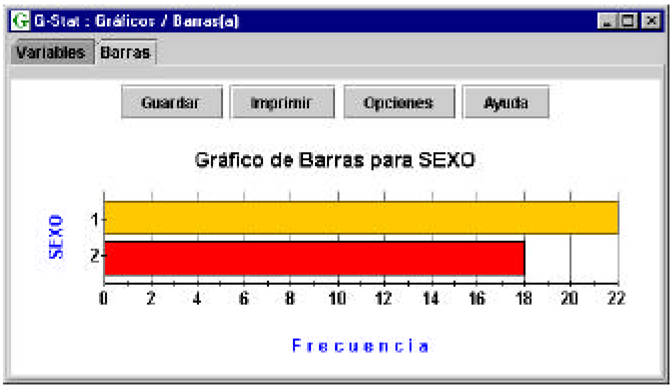


Gráfico de barras de la variable Sexo del *Ejemplo.*

Histograma (y)

Crea un histograma para una variable cuantitativa. El histograma, como paso previo, discretiza los valores de la variable en un número manejable de clases. La altura de cada bloque en el histograma depende del número de casos en cada clase. En un histograma se puede ver claramente cuál es la distribución de los datos. Normalmente, para el número de clases del histograma se toma la raíz cuadrada del número de casos.

Manejo del programa

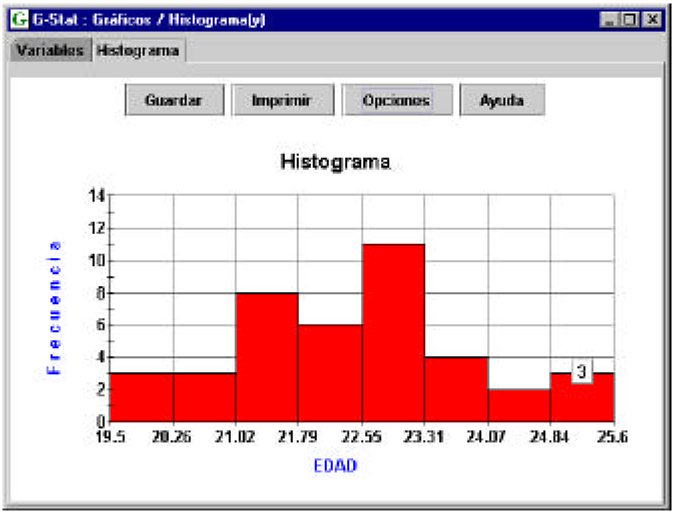
Identificar la variable a analizar. En la pestaña **Histograma** se encuentra la ventana con el histograma de la variable.

Opciones:

- La cabecera, los títulos de los ejes X e Y, el mínimo y máximo del eje X.

- Número de clases en el que se quiera discretizar la variable: Por defecto 5. El programa no admite más de 12 clases.

Obtener el histograma de la variable Edad con 8 clases.



Histograma de la variable Edad.

Cajas (y)

El diagrama de cajas es un gráfico que muestra la distribución de una variable cuantitativa, representando una serie de medidas de centralización. Su estructura está formada por una caja, figura rectangular, y dos segmentos horizontales situados a ambos lados de ésta.

Los bordes de la caja representan los cuartiles 1º y 3º, respectivamente, y la mediana corresponde a la línea central. Cuando la línea de la mediana se superpone con alguna línea de los cuartiles, no es posible distinguirla.

La media aparece señalada con un cuadrado gris y suele encontrarse próxima a la mediana. La distancia entre ambos valores, aporta información en cuanto a la simetría o asimetría de la variable. Cuando la variable es simétrica, media y mediana coinciden. La distancia entre estas dos medidas indica, asimismo, la posibilidad de valores extremos ya que la media es considerablemente sensible a ellos y la mediana no.

Los valores que estén situados a una distancia superior a 1.5 veces la distancia intercuartílica (diferencia entre el tercer y primer cuartil) son considerados “outliers” o valores extremos y están señalados en rojo. Los extremos de los segmentos corresponden al mínimo y al máximo de los valores sin considerar los valores extremos. En el caso de que no haya valores extremos, los segmentos son simplemente el mínimo y el máximo.

El nombre de la variable aparece en el gráfico. Este gráfico es autoescalable, pudiendo modificarse la relación entre longitud y anchura de la ventana mediante la posición del extremo inferior derecho.

Posicionando el cursor en cada punto aparecen las coordenadas.

Manejo del programa

Identificar la variable a analizar. En la pestaña **Cajas** aparece el gráfico de cajas de la variable seleccionada.

Opciones:

- La cabecera, orientación del gráfico, título, mínimo, máximo e incremento del eje X.

Series Temporales (y)

Crea una serie temporal teórica basada en los datos de una variable tiempo-dependiente. Se representa los valores de la variable seleccionada en el eje Y. Se asume que los valores de la variable representada en el gráfico están ordenados y que éstos están igualmente espaciados en el tiempo, representado en el eje X.

La serie predicha se representa como Y^. En la gráfica se presenta el ECM (Error Cuadrático Medio), que se interpreta como una medida del error en la predicción. A menor valor del ECM mejor es la serie estimada y más se acerca la modelización a los datos reales. El ECM viene dador por

$$ECM = \frac{1}{n} \sum e_t^2 \quad \text{con} \quad e_t = y_t - \hat{y}_t$$

Esta serie Y^ puede ser modelizada mediante media móvil, alisado exponencial y alisado exponencial doble. No calcula el gráfico cuando la variable analizada tiene menos de cuatro valores.

Media móvil

La media móvil de parámetro s se define como

$$M_t = \frac{y_t + y_{t-1} + \ldots + y_{t-s+1}}{s}$$

$$\hat{y}_t = M_t$$

Alisado exponencial

El alisado exponencial de parámetro alfa se define como

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t$$

$$\hat{y}_1 = y_1$$

Alisado exponencial doble de Holt-Winters

El alisado exponencial doble de parámetros alfa y beta se define como

$$M_t = \alpha y_t + (1-\alpha)(M_{t-1} + b_{t-1}) \text{ con } M_1 = y_1$$

$$b_t = \beta(M_t - M_{t-1}) + (1-\beta)b_{t-1} \text{ con } b_1 = 0$$

$$\hat{y}_{t+h} = M_t + b_t h$$

con:

- Alfa: Determina el peso dado a las observaciones ultimas en relación a las observaciones anteriores. Se define entre 0 y 1. Valores cercanos a 0 implican que las observaciones anteriores cuentan tanto como las más recientes. Valores cercanos a 1 indican lo contrario. Un valor de alfa igual a 1 indica que solo cuenta la última observación

- Beta: Determina el peso dado a las observaciones ultimas en relación a las observaciones anteriores en la estimación de la tendencia de la serie. Se define entre 0 y 1. Valores cercanos a 1 aumentan el peso relativo de las observaciones más recientes.

Manejo del programa

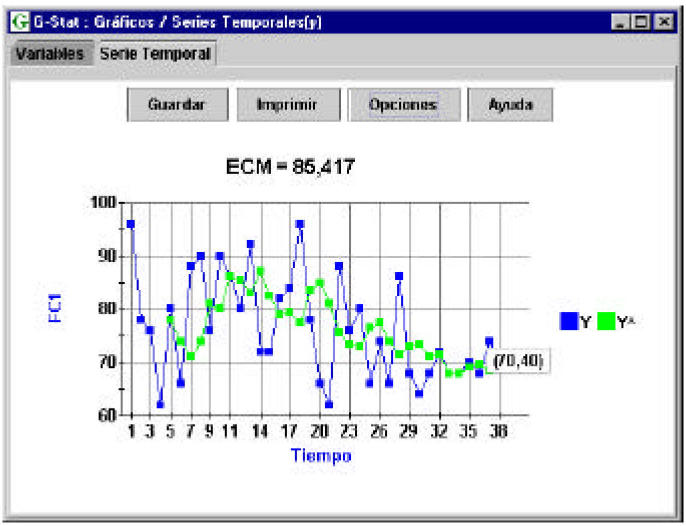
Identificar la variable a analizar. En la pestaña **Serie Temporal** se encuentran dos serie de datos, la original y la transformada.

Opciones:

- La cabecera, título, mínimo, máximo e incremento del eje Y.

- Tipo de serie temporal: Media móvil de parámetro S, alisado exponencial de parámetro alfa y alisado exponencial de doble parámetro alfa y beta.

Modelizar la variable FC1 con una serie temporal de medias móviles de parámetro S=4.



Serie temporal con media móvil S=4 de la variable FC1.

Barras (a|b)

Genera un gráfico de barras para dos variables cualitativas. Es una representación gráfica de las frecuencias de las celdas en tablas de frecuencias conjuntas de dos variables cualitativas / discretas. Se identifican las dos variables en la ventana de diálogo correspondiente como variables “respuesta” y “explicativa”, ambas cualitativas. La variable explicativa corresponde a la variable que forma los grupos. La variable respuesta es la que formará los bloques de frecuencias. Las alturas de los bloques corresponden a las frecuencias de cada combinación de niveles en las dos variables.

Manejo del programa

Se identifica la *Variable respuesta* y la *Variable explicativa*, amabas cualitativas. En la pestaña **Barras**, se encuentra el gráfico de barras bidimensional.

Opciones:

- La cabecera, orientación del gráfico, el título del eje X, su escala (absoluta/frecuencias, relativa/porcentajes), mínimo, máximo e incremento.

Cajas (a|y)

Crea un gráfico de cajas para una variable cuantitativa estratificada por una variable cualitativa. Es la representación gráfica de las distribuciones de diferentes submuestras de una variable cuantitativa. Es necesario identificar dos variables: una explicativa cualitativa / discreta (a) que es la que formará las submuestras, y otra respuesta cuantitativa (y) de la que se analizan los datos. Para cada nivel de la variable que forma las submuestras se presenta un diagrama de cajas.

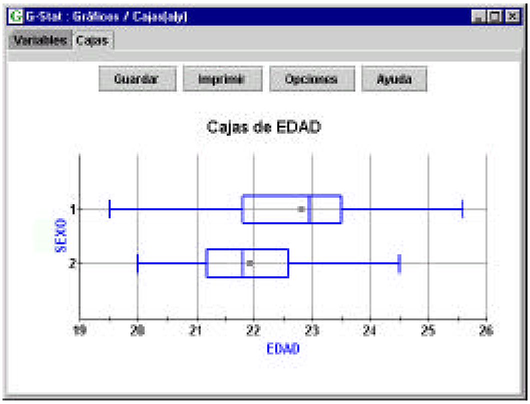
Manejo del programa

Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* cualitativa. En la pestaña **Cajas** se encuentra el gráfico correspondiente.

Opciones:

- La cabecera, orientación del gráfico, título, mínimo, máximo e incremento del eje X.

Obtener las cajas de Edad según el Sexo de los sujetos.



Gráficos de Cajas (a|y) de la variable Edad por Sexo.

Bloques de Medias y Desviaciones (a|y)

Crea un gráfico de bloques para las medias y segmentos para las desviaciones típicas de los valores de diferentes grupos. La variable respuesta debe ser cuantitativa y la variable formadora de grupos cualitativa.

Manejo del programa

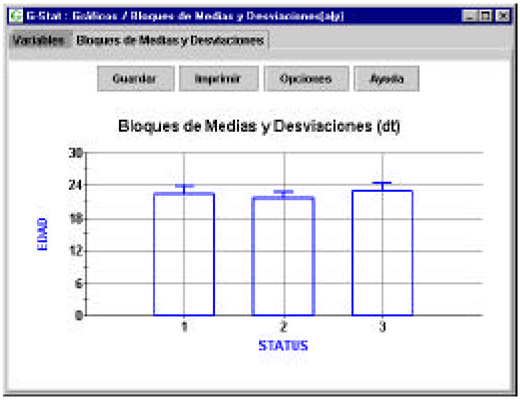
Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* cualitativa. En la pestaña **Bloques de Medias y Desviaciones** se encuentra el gráfico.

Opciones:

- La cabecera, título de los ejes X e Y, mínimo, máximo e incremento del eje Y.

- Desviaciones típicas o error estándar.

Representar las medias y desviaciones típicas de la variable Edad por grupos de Status.



Bloques de medias y desviaciones típicas de Edad por Status.

Dispersión (x|y)

Representación de la nube de puntos en ejes cartesianos de dos variables cuantitativas.

Manejo del programa

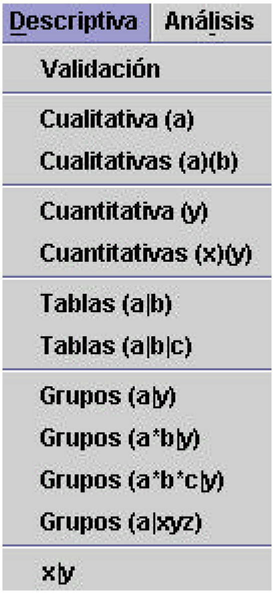
Se introduce la *Variable x* y la *Variable y*. En la pestaña **Dispersión** se encuentra el gráfico de dispersión de ambas variables.

Opciones:

- La cabecera, títulos, mínimo, máximo e incremento de los ejes X e Y.

Para la obtención de la recta de regresión y de sus límites confidenciales, acceder a los menús “Descriptiva / x|y” y “Análisis/ x|y”.

Menú Descriptiva



Activar la opción **Descriptiva** del menú principal o mediante Alt+D. Este menú, contiene las opciones necesarias para la realización de análisis descriptivos. Están separadas en seis grupos: validación, descriptiva de variables cualitativas, descriptiva de variables cuantitativas, tablas, descriptiva por grupos y relación entre dos variables cuantitativas. Los códigos (a) o (b) indican variables cualitativas y los códigos (y), (x) o (z) indican variables cuantitativas.

El asterisco implica que los resultados se estratificarán para todas las posibles combinaciones de categorías o niveles de las variables explicativas.

Validación

Se presenta un resumen básico de las variables que componen el fichero de trabajo en términos de número de casos y de variables así como el número de casos válidos y casos numéricos, mínimo y máximo. Esta descriptiva sirve de comprobación para detectar posibles errores en la entrada de datos. Para una descriptiva más precisa conviene usar el menú Descriptiva / Cualitativas (a)(b) para variables cualitativas y el menú Descriptiva / Cuantitativas (x)(y) para variables cuantitativas.

Realizar la validación de la base de datos del Ejemplo

Ventana de resultado al seleccionar validación en la base de datos del ejemplo.

Validación de Variables

Número de Casos: 40

Variable	Casos Válidos	Casos Numéricos	Mínimo	Máximo
IB	40	40	1.0	40.0
SEXO	40	40	1.0	2.0
FUMADOR	40	40	1.0	2.0
EDAD	40	40	19.5	25.6
FC1	40	40	62.0	96.0
FC2	40	40	112.0	165.0
FC2FC1	40	40	42.0	82.0
STATUS	40	40	1.0	3.0
FARMACO	40	40	1.0	2.0

Cualitativa (a)

Realiza la descriptiva para una variable cualitativa o discreta (a).

Manejo del programa

Variables: Se identifica la *Variable* cualitativa o discreta a analizar.

Frecuencias: Los principales estadísticos descriptivos para variables cualitativas son: las frecuencias absolutas y las frecuencias relativas. Para cada categoría de la variable cualitativa se muestra el número de individuos que pertenecen a ella (frecuencias absolutas), así como el porcentaje respecto al total de individuos (frecuencias relativas).

Barras: Gráfico de barras para una variable cualitativa. Ver menú Gráficos.

Cualitativas (a)(b)

Realiza la descriptiva para una o varias variables cualitativas o discretas.

Manejo del programa

Variables: Se identifican las *Variables* cualitativas o discretas.

Frecuencias: Permite obtener una descriptiva en términos de frecuencias para varias variables cualitativas a la vez. En cada variable, para cada categoría se muestra el número de individuos que pertenecen a ella (frecuencias absolutas), así como el porcentaje respecto al total de individuos (frecuencias relativas).

Cuantitativa (y)

Realiza la estadística descriptiva para una variable cuantitativa. Los principales estadísticos descriptivos para una variable cuantitativa son: media aritmética, mediana, moda, media geométrica, varianza, desviación típica, error estándar de la media, mínimo, máximo, rango o amplitud, cuartiles, rango intercuartílico, coeficiente de asimetría, coeficiente de asimetría estandarizada, coeficiente de curtosis, coeficiente de curtosis estandarizada y coeficiente de variación.

La **media aritmética** se calcula como la suma de los valores de las observaciones dividido por el tamaño muestral (n):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

La **mediana** (med) se calcula ordenando los datos de menor a mayor y tomando el valor del medio que es el que deja un 50% de observaciones a su izquierda y un 50% a su derecha. En el caso de que el número de observaciones sea par, la mediana se calcula como la semisuma de los dos valores centrales.

La **moda** es el valor que más se repite. Tiene sentido en variables con pocos niveles.

La **media geométrica** (mg) se calcula como la raíz enésima del producto de los valores de las observaciones, con

$$mg = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Otra expresión para su cálculo es evaluar la exponencial de la media aritmética de los logaritmos neperianos de las observaciones. Cuando existan valores negativos, el programa devuelve el valor “No Aplicable” para la media geométrica.

La **varianza** se calcula como

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2$$

La **desviación típica** o desviación estándar s se calcula como la raíz cuadrada positiva de la varianza de forma que

$$s = +\sqrt{s^2} = +\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$

El **error estándar** de la media se utiliza para estimar una media poblacional mediante intervalos de confianza. Su expresión es la desviación típica dividida entre la raíz cuadrada del tamaño muestral.

El **mínimo** es el menor valor observado, el **máximo** es el mayor valor observado y la **amplitud** (a veces llamada rango) es la diferencia entre el máximo y el mínimo.

Existen tres **cuartiles**: cuartil inferior, cuartil medio y cuartil superior. El cuartil inferior se calcula ordenando los datos de menor a mayor y tomando el valor que deja un 25% de observaciones a su izquierda y un 75% a su derecha. El cuartil medio es la mediana. El cuartil superior es aquel valor, que en los datos ordenados, deja un 75% a su izquierda y un 25% a su derecha. El **rango intercuartílico** es la diferencia entre el cuartil superior y el inferior.

El **coeficiente de asimetría** se calcula como

$$\frac{n}{(n-1)(n-2)} \frac{1}{s^3} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^3$$

El **coeficiente de asimetría estandarizada** viene dado por

$$\frac{\text{asimetría}}{\sqrt{\frac{6}{n}}}$$

El **coeficiente de curtosis** (apuntamiento) se calcula como

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{1}{s^4}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4-3\frac{(n-1)(n-1)}{(n-2)(n-3)}$$

El **coeficiente de curtosis estandarizada** viene dado por

$$\frac{\text{curtosis}}{\sqrt{\frac{24}{n}}}$$

El **coeficiente de variación** se calcula como

$$\frac{s}{\bar{x}}$$

Los **percentiles** son aquellos valores que dejan un p% de observaciones a un lado de su valor y un (1-p)% al otro, siendo p un número cualquiera entre cero y uno. Sea n el tamaño muestral, el percentil p se calcula como:

$$\frac{1}{2}\left(x_{(i)}+x_{(i+1)}\right) \text{ si } f = 0$$

$$x_{(i+1)} \text{ si } f > 0$$

siendo i la parte entera de n·p y f la parte fraccional de n·p, y donde (i) indica el valor ordenado de los valores de la variable x de menor a mayor que ocupa la posición i-ésima.

Por ejemplo si n=40 y p=25%, se tiene que i=10 y f=0. Si n=39 y p=50%, se tiene que i=19 y f=0.5.

Observar que si n es par, la mediana (percentil 50%) se calcula como la semisuma de los dos valores centrales y que si n es impar como el valor que deja a la izquierda y a la derecha el mismo número de valores.

Ejemplos de percentiles son los deciles y los cuartiles. Los deciles son los percentiles en donde p=0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Los cuartiles son los percentiles en donde p=0.25, 0.50, 0.75. Al cuartil p=0.25 se le conoce como cuartil inferior o primer cuartil Q1, al cuartil p=0.75 como cuartil superior o tercer cuartil Q3. La mediana es el segundo cuartil Q2.

Manejo del programa

Variables: Se identifica la *Variable* cuantitativa.

Estadísticos: Presenta los estadísticos seleccionados para esta variable.

Opciones:

- Seleccionar los estadísticos a visualizar: por defecto aparecen todos.

Cajas: Gráfico de Cajas de la variable cuantitativa. Ver menú Gráficos.

Clases: Con esta opción se discretiza la variable cuantitativa en distintos tramos o intervalos. Para cada uno de ellos se suministra información en términos de frecuencias absolutas y relativas. Las frecuencias absolutas indican el número de individuos para cada intervalo, las relativas la proporción respecto al total.

Opciones:

- Número de clases: Por defecto el programa considera 5 clases.

Histograma: Histograma para una variable cuantitativa. Ver menú Gráficos.

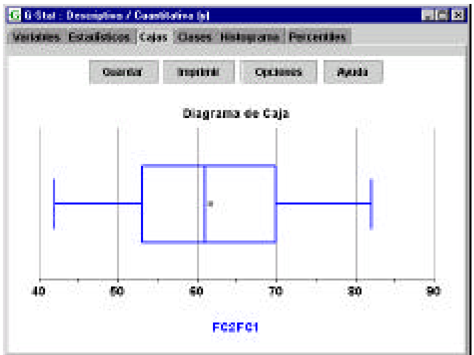
Percentiles: Presentan como opción por defecto, los percentiles del 1%, 5%, 10%, 25%, 75%, 90%, 95% y 99%.

Realizar un análisis descriptivo de la variable FC2FC1.

Estadísticos de la variable FC2FC1 con la opción Cuantitativa (y).

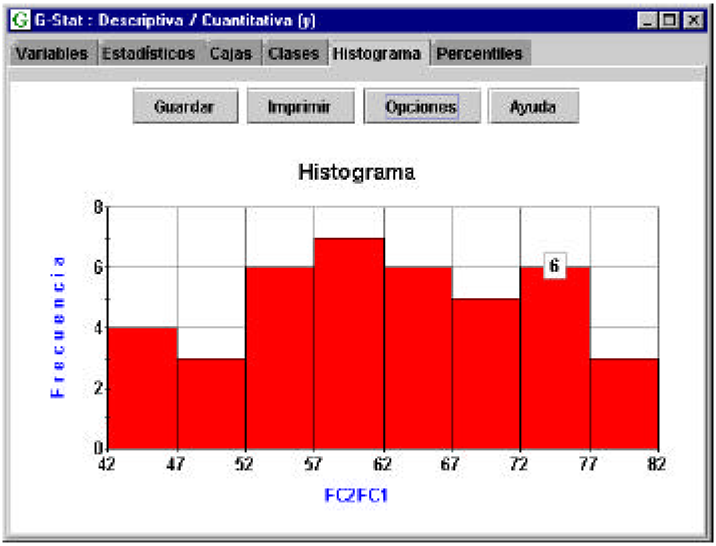
```
Estadísticos para la variable FC2FC1
======================================================================
---------------------------------------
 Estadístico                 FC2FC1
---------------------------------------
 N                           40
 Media                       61.7500
 Mediana                     61.0000
 Moda                        52.0000
 Media Geométrica            60.7718
 Varianza                    120.5000
 Desviación Típica           10.9772
 E.E. de la Media (*)        1.7357
 Mínimo                      42.0000
 Máximo                      82.0000
```

```
 Rango                          40.0000
 Cuartil Inferior               53.0000
 Cuartil Superior               70.0000
 Rango Intercuartílico          17.0000
 Asimetría                      -0.0168
 Asimetría  Estandarizada       -0.0433
 Curtosis                       -0.8432
 Curtosis Estandarizada         -1.0885
 Coeficiente de Variación       17.7769
---------------------------------------

(*) Usar con propósito de estimación para el I.C. de la media
```



Cajas de la variable FC2FC1 de la opción Cuantitativa (y).

```
Clases de la variable FC2FC1
======================================================================

Número de Casos: 40

---------------------------------------------------------------------
        Límite     Límite     Frec.      Frec. Abs.Frec.     Frec. Rela.
 Clases Inferior   Superior   Absoluta   Acumulada Relativa Acumulada
---------------------------------------------------------------------
 1      42.00      50.00      6          6         0.15     0.15
 2      50.00      58.00      7          13        0.17     0.32
 3      58.00      66.00      11         24        0.28     0.60
 4      66.00      74.00      9          33        0.23     0.82
 5      74.00      82.00      7          40        0.17     1.00
---------------------------------------------------------------------
```



Histograma de la opción Cuantitativa (y) para FC2FC1.

Percentiles de la variable FC2FC1.

```
Percentiles para la variable FC2FC1
==================================================================
Número de Casos: 40
Percentiles

 1.0%    42.00
 5.0%    43.00
 10.0%   47.00
 25.0%   53.00
 75.0%   70.00
 90.0%   76.00
 95.0%   80.00
 99.0%   82.00
```

Cuantitativas (x)(y)

Realiza la estadística descriptiva para varias variables cuantitativas.

Manejo del programa

Variables: Se identifican las *Variables* a analizar.

Estadísticos: Estadísticos de las variables seleccionadas. Se detallan en Cuantitativa (y).

Correlaciones: Matriz de correlaciones de Pearson y Spearman de las variables seleccionadas considerando los casos válidos. Ver Descriptiva / x|y / Modelo para las definiciones de dichos coeficientes y Análisis / x|y / Modelo para ver cómo se calculan sus significaciones.

Se desea calcular los coeficientes de correlación de Pearson de los datos del ejemplo para las variables FC1, FC2 y Edad.

Matriz de coeficientes de correlación de las variables FC1, FC2 y Edad.

```
Cuantitativas (x)(y). Correlaciones
==========================================================================
Variables         : FC1, FC2, EDAD
Número de Casos   : 40

r de Pearson
(Significación)

         FC1          FC2          EDAD
-------------------------------------------

 FC1     1.0000       0.5796       0.1848
                      (0.0001)     (0.2537)

 FC2     0.5796       1.0000       -0.6437
         (0.0001)                  (0.0001)

 EDAD    0.1848       -0.6437      1.0000
         (0.2537)     (0.0001)
-------------------------------------------
```

Tablas (a|b)

Contiene un submenú para variables cualitativas con datos no agrupados y agrupados.

Tablas (a|b) → Tablas

Presenta una tabla de contingencia para dos variables cualitativas o discretas y los estadísticos descriptivos asociados.

En las tablas de contingencia, se recoge la frecuencia absoluta del número de individuos para cada una de las posibles combinaciones de niveles de las dos variables. Estas frecuencias absolutas se pueden relativizar respecto al total de cada nivel en cada variable (porcentaje de filas y columnas) o respecto al total de individuos (porcentaje total).

Se muestran los principales estadísticos de asociación entre dos variables cualitativas: los estadísticos de asociación y los estadísticos con modelo.

La notación que se sigue es la de una matriz con r filas y c columnas donde:

	Var. en columnas=var X				
Var. en filas=var Y	Cat1	Cat2	...	Catc	Total
Niv1	n_{11}	n_{12}	...	n_{1c}	r_1
Niv2	n_{21}	n_{22}	...	n_{2c}	r_2
...	...	...	...	...	...
Nivr	n_{r1}	n_{r2}	...	n_{rc}	r_r
Total	c_1	c_2	...	c_c	n

Los estadísticos de asociación que se tratan son: V de Cramer, coeficiente de contingencia C, lambda simétrica λ_{sim}, coeficiente de incertidumbre simétrico U_{sim}, Gamma de Goodman-Kruskal γ_{G_k} , Tau-b de Kendall, Tau-c de Stuart y D_{sim} de Somer simétrico. Los estadísticos con modelo que se tratan son: lambda asimétrica λ_{asim}, coeficiente de incertidumbre asimétrico U_{asim} y D_{asim} de Somer asimétrico.

En el caso de que haya al menos una variable cualitativa sólo se muestran los siguientes estadísticos: V de Cramer, coeficiente de contingencia C, lambda simétrica λ_{sim}, coeficiente de incertidumbre simétrico U_{sim}, lambda asimétrica λ_{asim} y coeficiente de incertidumbre asimétrico U_{asim}. En el caso de que las dos variables sean cuantitativas se muestran todos los estadísticos considerados.

V de Cramer

$$V = +\sqrt{\frac{\chi^2 / n}{\text{mín}\{r-1,\ c-1\}}}$$

donde χ^2 es el valor del estadístico de contraste Chi-Cuadrado para una tabla de dimensiones r x c (ver Análisis / Tablas (a|b) / Chi-Cuadrado / Chi-Cuadrado).

Coeficiente de contingencia C

$$C = +\sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Lambda simétrica $l_{\mathbf{sim}}$

$$\lambda_{sim} = \frac{\sum_i \underset{j}{\text{máx}}\, n_{ij} + \sum_j \underset{i}{\text{máx}}\, n_{ij} - \underset{j}{\text{máx}}\, c_j - \underset{i}{\text{máx}}\, r_i}{2n - \underset{j}{\text{máx}}\, c_j - \underset{i}{\text{máx}}\, r_i}$$

Coeficiente de incertidumbre simétrico $\mathbf{U_{sim}}$

$$U_{sim} = \frac{2[H(x) + H(y) - H(xy)]}{[H(x) + H(y)]} \quad \text{con}$$

$$H(x) = -\sum_{i=1}^{r} \frac{r_i}{n} \operatorname{Ln}\left(\frac{r_i}{n}\right)$$

$$H(y) = -\sum_{j=1}^{c} \frac{c_j}{n} \operatorname{Ln}\left(\frac{c_j}{n}\right)$$

$$H(xy) = -\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}}{n} \operatorname{Ln}\left(\frac{n_{ij}}{n}\right)$$

Gamma de Goodman-Kruskal γ_{Gk}

$$\gamma_{GK} = \frac{P' - Q'}{P' + Q'} \quad \text{con}$$

$$P' = \sum_i \sum_j n_{ij} A_{ij}$$

$$A_{ij} = \sum_{k>i}\sum_{l>j} n_{kl} + \sum_{k<i}\sum_{l<j} n_{kl}$$

$$Q' = \sum_{i}\sum_{j} n_{ij} D_{ij}$$

$$D_{ij} = \sum_{k>i}\sum_{l<j} n_{kl} + \sum_{k<i}\sum_{l>j} n_{kl}$$

Tau-b de Kendall τ_b

$$\tau_b = \frac{P' - Q'}{\sqrt{\left[n(n-1) - \sum_i r_i(r_i - 1)\right]\left[n(n-1) - \sum_j c_j(c_j - 1)\right]}}$$

Tau-c de Stuart τ_c

$$\tau_c = \frac{P' - Q'}{n^2 \frac{m-1}{m}} \quad \text{con } m = \text{mín }\{r, c\}$$

Coeficiente D de Somer simétrico

$$D_{sim} = 2 \cdot \frac{P' - Q'}{w_r + w_c} \quad \text{con}$$

$$w_r = n^2 - \sum_i r_i^2$$

$$w_c = n^2 - \sum_j c_j^2$$

Lambda asimétrica $\lambda_{\mathbf{asim}}$

Suponiendo que la variable fila es la variable respuesta, es

$$\lambda_{asim} = \frac{\sum_j \underset{i}{\text{máx}}\, n_{ij} - \underset{i}{\text{máx}}\, r_i}{n - \underset{i}{\text{máx}}\, r_i}$$

y suponiendo que la variable columna es la variable respuesta, es

$$\lambda_{asim} = \frac{\sum_i \max_j n_{ij} - \max_j c_j}{n - \max_j c_j}$$

Coeficiente de incertidumbre asimétrico U_{asim}

Suponiendo que la variable fila es la variable respuesta, es

$$U_{asim} = \frac{H(x) + H(y) - H(xy)}{H(x)}$$

y suponiendo que la variable columna es la variable respuesta, es

$$U_{asim} = \frac{H(x) + H(y) - H(xy)}{H(y)}$$

Coeficiente D de Somer asimétrico

Suponiendo que la variable fila es la variable respuesta, es

$$D_{asim} = \frac{P' - Q'}{w_c}$$

y suponiendo que la variable columna es la variable respuesta, es

$$D_{asim} = \frac{P' - Q'}{w_r}$$

Manejo del programa

Variables: Se identifica la *Variable* que aparecerá en filas y la *Variable* que aparecerá en columnas.

Tablas: Se muestra la tabla de contingencia de dos variables cualitativas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Estadísticos: Se muestran los principales estadísticos de asociación entre dos variables cualitativas:

V de Cramer
Coeficiente de Contingencia
Lambda simétrica
Coeficiente de incertidumbre simétrico
Gamma de Goodman-Kruskal
Tau-b de Kendall
Tau-c de Stuart
D de Somer simétrico

También se presentan los siguientes estadísticos descriptivos cuando se asume un modelo

Lambda asimétrica
Coeficiente de incertidumbre asimétrico
D de Somer asimétrico

Barras: Representación gráfica de las frecuencias de las celdas en tablas de frecuencias conjuntas de dos variables cualitativas /discretas. Se identifican las dos variables en la ventana de diálogo correspondiente como variables “fila” y “columna”. La variable fila corresponde a la variable que forma los grupos. La variable columna es la que formará los bloques de frecuencias. Las alturas de los bloques corresponden a las frecuencias de cada combinación de niveles en las dos variables.

Opciones:

- La cabecera, orientación del gráfico, el título del eje X, su escala (absoluta/frecuencias, relativa/porcentajes), mínimo, máximo e incremento.

Tablas (a|b) → Tablas. Datos agrupados

A partir de datos agrupados se calculan los estadísticos descriptivos asociados para dos variables cualitativas o discretas. Los datos agrupados se introducen directamente en una cuadrícula en forma de tabla de r filas y c columnas. Los fundamentos teóricos y la formulación son idénticos a los presentados en la opción de análisis anterior con datos a partir de un fichero.

Manejo del programa

Datos Agrupados: La cuadrícula permite la entrada directa de las frecuencias. Se puede definir el número de categorías o niveles de las dos variables mediante el número de filas y columnas. Por defecto aparecen unos valores que deben ser sustituidos por los datos del usuario.

No dejar filas o columnas con valores faltantes o con todos los valores cero. El Botón "Crear Tabla" prepara la estructura de la tabla ajustada al número de filas y columnas definido. La tabla no admite valores negativos, decimales o alfanuméricos.

Tablas: Se muestra la tabla de contingencia de dos variables cualitativas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Estadísticos: Se muestran los principales estadísticos de asociación entre dos variables cualitativas:

V de Cramer
Coeficiente de Contingencia
Lambda simétrica
Coeficiente de incertidumbre simétrico
Gamma de Goodman-Kruskal
Tau-b de Kendall
Tau-c de Stuart
D de Somer simétrico

También se presentan los siguientes estadísticos descriptivos cuando se asume un modelo

Lambda asimétrica
Coeficiente de incertidumbre asimétrico
D de Somer asimétrico

Barras: Representación gráfica de las frecuencias de las celdas en tablas de frecuencias conjuntas de dos variables cualitativas /discretas.

Se identifican las dos variables en la ventana de diálogo correspondiente como variables "fila" y "columna". La variable fila corresponde a la variable que forma los grupos. La variable columna es la que formará los bloques de frecuencias. Las alturas de los bloques corresponden a las frecuencias de cada combinación de niveles en las dos variables.

Opciones:

- La cabecera, orientación del gráfico, el título del eje X, su escala (absoluta/frecuencias, relativa/porcentajes), mínimo, máximo e incremento.

Tablas (a|b|c)

Presenta tablas de frecuencias para dos variables cualitativas, estratificadas por una tercera variable cualitativa que forma las capas.

Permite realizar un análisis estratificado mediante varias tablas de contingencia de las variables a y b para cada uno de los valores de la variable c (la que forma las capas). En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías. Adicionalmente se presentan los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total).

Manejo del programa

Variables: Se identifican las dos variables que formarán las tablas, identificando la *Variable* que aparecerá en filas y la *Variable* que aparecerá en columnas. Adicionalmente se identifica una tercera *Variable* cualitativa formadora de las capas o estratos. Sólo aparecen los registros completos para estas tres variables.

Tablas: Aparece el número total de casos válidos no faltantes y varias tablas cruzadas de las variables seleccionadas en filas y columnas, para cada nivel de la variable en capas. En las tablas aparecen las frecuencias absolutas y los porcentajes por filas, columnas o por totales. El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Grupos (a|y)

Realiza una descriptiva para una variable cuantitativa, estratificada por otra variable cualitativa o discreta.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* cualitativa. Sólo se utilizan los registros completos para estas variables.

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

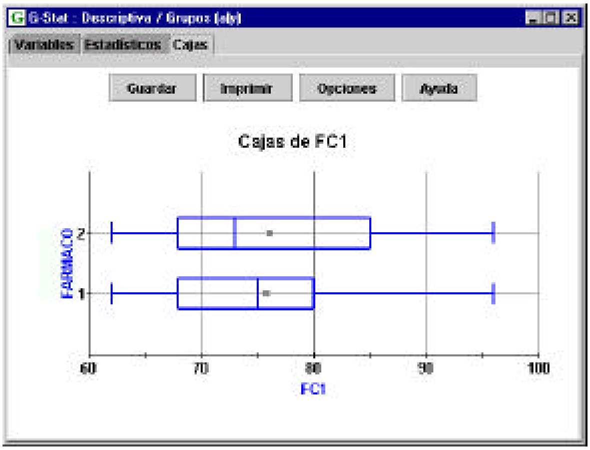
Cajas: Gráfico de Cajas de la variable respuesta para los distintos valores de la variable explicativa. Ver menú Gráficos.

Descriptiva de la variable FC1 según el tipo de Fármaco que se ha administrado.

Estadísticos de la variable FC1 por Farmaco mediante la opción Grupos (a|y).

Estadísticos para la variable FC1 por FARMACO

Grupo	1	2
N	20.0	20.0
Media	75.8000	76.1000
Mediana	75.0000	73.0000
Varianza	90.0632	96.6211
Desviación Típica	9.4902	9.8296
Mínimo	62.0000	62.0000
Máximo	96.0000	96.0000
Cuartil Inferior	68.0000	68.0000
Cuartil Superior	80.0000	85.0000
Asimetría Estandarizada	1.3675	0.7484
Curtosis Estandarizada	-0.2461	-0.8821
Coeficiente de Variación	12.5200	12.9167



Cajas de la variable FC1 estratificada por la

variable Farmaco mediante la opción Grupos (a|y).

Grupos (a*b|y)

Descriptiva para una variable cuantitativa y estratificada por todas las posibles combinaciones de niveles de las variables cualitativas a y b.

Manejo del programa

Variables: Se identifica la variable respuesta cuantitativa, y las variables explicativas cualitativas a y b formadoras de grupos. Sólo se utilizan los registros completos para estas variables.

Estadísticos: Estadísticos de la variable seleccionada para cada posible categoría de las variables cualitativas combinadas. Se detallan en Cuantitativa (y).

Descriptiva de la variable Edad según las variables Sexo y Fumador.

Resultados de la descriptiva de una variable cuantitativa por subgrupos formados por dos variables cualitativas.

Grupos (a*b|y) Estadísticos

Variable Respuesta: EDAD
Variable(s) Explicativa(s): SEXO, FUMADOR
Número de Casos: 40

SEXO	N	Media	Desviación Típica	Mínimo	Máximo
2	18	21.9333	1.1555	20.0000	24.5000
1	22	22.8091	1.5486	19.5000	25.6000
Total	40	22.4150	1.4380	19.5000	25.6000

FUMADOR	N	Media	Desviación Típica	Mínimo	Máximo
2	24	22.5833	1.3871	19.5000	25.6000
1	16	22.1625	1.5209	20.0000	25.4000
Total	40	22.4150	1.4380	19.5000	25.6000

SEXO FUMADOR	N	Media	Desviación Típica	Mínimo	Máximo
2,2	12	22.3583	1.0211	21.1000	24.5000
2,1	6	21.0833	0.9725	20.0000	22.5000
1,2	12	22.8083	1.6941	19.5000	25.6000
1,1	10	22.8100	1.4449	20.5000	25.4000
Total	40	22.4150	1.4380	19.5000	25.6000

Grupos (a*b*c|y)

Realiza una descriptiva para una variable cuantitativa (y) estratificada por todas las posibles combinaciones de niveles de las variables cualitativas (a), (b) y (c).

Manejo del programa

Variables: Se identifican la *Variable respuesta* cuantitativa y las *Variables explicativas* cualitativas que forman los grupos. Sólo se utilizan los registros completos para estas variables.

Estadísticos: Estadísticos de la variable respuesta para cada posible combinación de las categorías de las tres variables cualitativas. Se detallan en Cuantitativa (y).

Grupos (a|xyz)

Presenta la descriptiva de diferentes variables cuantitativas (x), (y), (z) estratificadas por una variable cualitativa o discreta formadora de los grupos (a).

Manejo del programa

Variables: Se identifica la *Variable explicativa* cualitativa y las *Variables respuesta* cuantitativas.

Estadísticos: Estadísticos de las variables respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Descriptiva de Edad, FC1, FC2 para los diferentes grupos de Fármacos.

Resultados de la descriptiva de varias variables cuantitativas por subgrupos formados por una variable cualitativa.

```
Grupos (a|xyz). Estadísticos
======================================================================

Variable que forma los grupos : FARMACO
Variable(s)                   : EDAD, FC1, FC2
Número de Casos:                40

Nota: Sólo se consideran los registros (casos) con información
completa en todas las variables analizadas

Variable=EDAD
                                            Desviación
 FARMACO   N        Media      Mediana     Típica         Mínimo     Máximo
---------------------------------------------------------------------------
 1         20       22.6000    22.8000     1.6588         19.5000    25.6000
 2         20       22.2300    22.1500     1.1921         20.0000    24.9000
---------------------------------------------------------------------------
 Total     40       22.4150    22.5500     1.4380         19.5000    25.6000

Variable=FC1
                                            Desviación
 FARMACO   N        Media      Mediana     Típica         Mínimo     Máximo
---------------------------------------------------------------------------
 1         20       75.8000    75.0000     9.4902         62.0000    96.0000
 2         20       76.1000    73.0000     9.8296         62.0000    96.0000
---------------------------------------------------------------------------
 Total     40       75.9500    74.0000     9.5379         62.0000    96.0000

Variable=FC2
                                             Desviación
 FARMACO   N        Media      Mediana      Típica        Mínimo     Máximo
---------------------------------------------------------------------------
 1         20       136.9500   137.0000     13.3435       116.0000   158.0000
 2         20       138.4500   137.0000     13.5199       112.0000   165.0000
---------------------------------------------------------------------------
 Total     40       137.7000    137.0000     13.2804        112.0000 165.0000
```

x|y

Realiza una descriptiva bivariante de dos variables cuantitativas. El modelo que se asume es:

$$y = \beta_0 + x\beta_1 + \varepsilon$$

Los coeficientes β_0 y β_1 se estiman por b_0 (ordenada en el origen) y por b_1 (pendiente) a través del método de mínimos cuadrados:

$$b_1 = \frac{s_{xy}^2}{s_x^2}\ ,\ \ b_0 = \overline{y} - b_1\overline{x}$$

donde

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ ,\ \ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\ ,\ \ s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

Por tanto, la ecuación resultante dada por el modelo de regresión lineal simple es

$$\hat{y} = b_0 + b_1 x$$

donde ^y indica el valor que predice el modelo para la variable respuesta a partir de la información de la variable explicativa.

Los residuos se calculan como la diferencia que hay entre la variable respuesta original y la que predice el modelo, es decir:

$$\text{residuos} = e = y - \hat{y}$$

La desviación típica residual se calcula como

$$s_e = \sqrt{\frac{1}{n-2}\sum\left(e - \overline{e}\right)^2}$$

Se puede demostrar que la media de los residuales es cero, con lo que

$$s_e = \sqrt{\frac{1}{n-2}\sum e^2}$$

Se divide por n - 2 en lugar de por n – 1 para obtener una estimación mejor de la desviación típica residual poblacional.

Los errores estándar (EE) de los coeficientes del modelo se calculan a partir de la desviación típica residual, de forma que

$$EE\left(b_0\right) = s_e\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}$$

$$EE(b_1) = s_e \frac{1}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}$$

Una medida global de bondad del modelo es el coeficiente de determinación R2 (“R-Cuadrado”). El coeficiente de determinación se calcula como

$$R^2 = \frac{SCM}{SCT}$$

donde SCM es la suma de cuadrados del modelo y SCT la suma de cuadrados total (ver expresiones en Análisis / x|y / Regresión Lineal Simple / Anova). R2 en tanto por ciento representa el porcentaje de información que explica el modelo. El coeficiente de determinación también se puede calcular como el cuadrado del coeficiente de correlación de Pearson.

El coeficiente de correlación de Pearson está comprendido entre –1 y +1, es adimensional y está íntimamente relacionado con la desviación típica residual. Conviene señalar que correlación implica asociación lineal, no implica que no haya otro tipo de asociación no lineal (como por ejemplo curvilínea o senoidal). Por otra parte correlación no implica causalidad.

El coeficiente de correlación r de Pearson está dado por

$$r = \frac{s_{xy}^2}{s_x s_y}$$

El coeficiente de correlación de Spearman es el análogo no paramétrico al coeficiente de correlación de Pearson, ya que utiliza los rangos de las variables y puede ser utilizado para variables ordinales o incluso dicotómicas o para variables cuantitativas con muestras pequeñas. El coeficiente de correlación de Pearson requiere normalidad en las variables.

Para calcular el coeficiente de correlación de Spearman entre dos variables Var1 y Var2, se calculan los rangos de los valores de éstas, a los que se denota por: R_i(Var1) y R_i(Var2), siendo R_i(Var1) los rangos de Var1 asociados al individuo i y R_i(Var2) los rangos de Var2 asociados al individuo i. A continuación, se realizan los siguientes cálculos intermedios:

$$D = \sum_{i=1}^{n}\left(R_i(\text{Var1}) - R_i(\text{Var2})\right)^2$$

$$T_x = \sum_{\text{empates en Var1}}\left(\text{nº empates}^3 - \text{nº empates}\right)$$

$$T_y = \sum_{\text{empates en Var2}}\left(\text{nº empates}^3 - \text{nº empates}\right)$$

$$A = \frac{n^3 - n - T_x}{12}, \quad B = \frac{n^3 - n - T_y}{12}$$

A partir de los coeficientes calculados con anterioridad, se calcula el coeficiente de correlación r_s de Spearman dado por

$$r_S = \frac{A + B - D}{2\sqrt{AB}}$$

Se puede demostrar que si se calcula el coeficiente de correlación de Pearson sobre las variables R_i(Var1) y R_i(Var2) se llega al mismo resultado.

Manejo del programa

Variables: Se identifican las dos *Variables* cuantitativas X e Y.

Estadísticos: Estadísticos de las variables seleccionadas. Se detallan en Cuantitativa (y).

Modelo: Se presentan los coeficientes de la recta de regresión lineal de mejor ajuste por el método de mínimos cuadrados. También se calculan los siguientes estadísticos: r de Pearson, r cuadrado, Desviación Típica de Residuos y Rho de Spearman.

Recta de Ajuste: Presenta la recta de regresión estimada por mínimos cuadrados.

Opciones:

- La cabecera, títulos, mínimo, máximo e incremento de los ejes X e Y.

Estudiar la relación lineal de las variable FC2 y Edad.

Estadísticos de la variable FC2 y Edad.

```
Regresión Lineal Simple. Estadísticos
======================================================================
Variable Y:       EDAD
Variable X:       FC2
Número de Casos:  40

------------------------------------------
```

```
 Variable             FC2          EDAD
------------------------------------------
 N                    40           40
 Media                137.7000     22.4150
 Mediana              137.0000     22.5500
 Desviación Típica    13.2804      1.4380
 Mínimo               112.0000     19.5000
 Máximo               165.0000     25.6000
 Rango                53.0000      6.1000
------------------------------------------
```

Modelo y Coeficientes de regresión y correlación de Edad por FC2.

```
Modelo de EDAD con FC2
====================================================================
Número de Casos:   40

Modelo: Lineal
--------------------------------
Ecuación: EDAD = 32.0126 - 0.0697 * FC2
--------------------------------
             Coef.     E.E.
--------------------------------
 Ordenada    32.0126   1.8595
 Pendiente   -0.0697   0.0134
--------------------------------
r de Pearson (coeficiente de correlación):  -0.6437
r cuadrado (coeficiente de determinación):  41.43%
Desviación Típica de los Residuos:          1.1149
Rho de Spearman:                            -0.6594
```

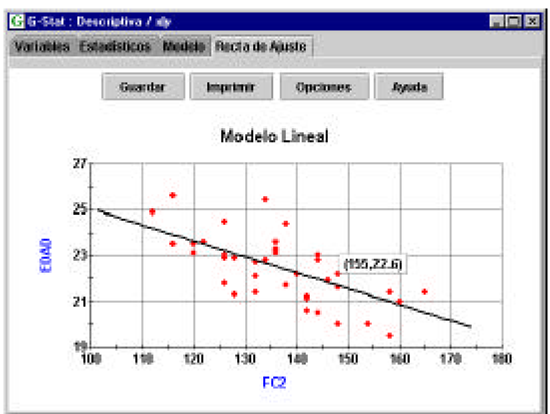


Gráfico de dispersión y recta de ajuste de Edad frente a FC2.

Menú Análisis

Activar la opción **Análisis** del menú principal o mediante Alt+L. Este menú contiene fundamentalmente las pruebas estadísticas univariantes y bivariantes, tanto para variables cualitativas como cuantitativas. Asimismo, se presenta una opción con los cálculos de las distribuciones teóricas de probabilidad directas e inversas. Los códigos (a) o (b) indican que el análisis ha sido diseñado para variables cualitativas y los códigos (y) o (x) para variables cuantitativas.

Distribuciones

A partir del valor de los estadísticos y de los grados de libertad, mediante esta opción, se pueden obtener las probabilidades asociadas a las siguientes distribuciones teóricas: Normal, t-Student, Chi-Cuadrado, F, Rango Estudentizado y Shapiro-Wilk. Inversamente se pueden obtener los valores de los estadísticos a partir de valores de probabilidad. Donde es apropiado se consideran las funciones unilaterales o bilaterales.

Consta del siguiente submenú: Normal, Normal Inversa, t-Student, t-Student Inversa, Chi-Cuadrado, Chi-Cuadrado Inversa, F, F Inversa, Rango Studentizado Inversa y Shapiro-Wilk.

Distribuciones → Normal

Para un valor z de una distribución normal se calcula el p-valor bilateral, el p-valor unilateral izquierdo y el p-valor unilateral derecho. Por ejemplo, para z=1.96 se tiene que:

$$p-\text{valor bilateral} = 2 \cdot \text{Prob}\left(N(0,1) \geq \left|1.96\right|\right) = 0.05$$

$$p-\text{valor unilateralizquierda} = \text{Prob}\left(N(0,1) \leq 1.96\right) = 0.9750$$

$$p-\text{valor unilateralderecha} = \text{Prob}\left(N(0,1) \geq 1.96\right) = 0.0250$$

Distribuciones → Normal Inversa

Para un valor alfa se calcula el correspondiente valor z bilateral y unilateral correspondiente a una distribución normal. Por ejemplo, para α=0.05 se tiene que:

$$z-\text{valor bilateral} = 1.96 \ \text{ que verifica } \ 2 \cdot \text{Prob}\left(N(0,1) \geq \left|1.96\right|\right) = 0.05$$

$$p-\text{valor unilateral} = 1.6449 \ \text{ que verifica } \ \text{Prob}\left(N(0,1) \geq 1.6449\right) = 0.05$$

Distribuciones → t-Student

Para un valor t de una distribución t-Student con g grados de libertad se calcula el p-valor bilateral, el p-valor unilateral izquierdo y el p-valor unilateral derecho. Por ejemplo, para t=-0.0750 y g=15 se tiene que:

$$p-\text{valor bilateral} = 2 \cdot \text{Prob}\left(t_g \geq \left|-0.0750\right|\right) = 0.9412$$

$$p-\text{valor unilateralizquierda} = \text{Prob}\left(t_g \leq -0.0750\right) = 0.4706$$

$$p-\text{valor unilateralderecha} = \text{Prob}\left(t_g \geq -0.0750\right) = 0.5294$$

Calcular la probabilidad asociada al valor del estadístico t-Student = 2.0421 para 20 grados de libertad.

Resultados de la opción t-Student.

```
t-Student
======================================================================

Para t = 2.0421 y gl = 20
p-valor bilateral = 0.0546
p-valor unilateral izquierda = 0.9727
p-valor unilateral derecha = 0.0273
```

Distribuciones → t-Student Inversa

Para un valor alfa se calcula el correspondiente valor t bilateral y unilateral correspondiente a una distribución t-Student con g grados de libertad. Por ejemplo, para $\alpha=0.05$ y g=15 se tiene que:

$$t-\text{valor bilateral} = 2.1314 \text{ que verifica } 2\cdot\text{Prob}\left(t_g \geq |2.1314|\right) = 0.05$$

$$t-\text{valor unilateral} = 1.7530 \text{ que verifica } \text{Prob}\left(t_g \geq 1.7530\right) = 0.05$$

Distribuciones → Chi-Cuadrado

Para un valor Chi-2 de una distribución Chi-Cuadrado con g grados de libertad se calcula el p-valor unilateral izquierdo y el p-valor unilateral derecho. Por ejemplo, para Chi-2=19.0228 y g=9 se tiene que:

$$p-\text{valor unilateral izquierda} = \text{Prob}\left(\chi^2_g \leq 19.0228\right) = 0.9750$$

$$p-\text{valor unilateral derecha} = \text{Prob}\left(\chi^2_g \geq 19.0228\right) = 0.0250$$

Distribuciones → Chi-Cuadrado Inversa

Para un valor alfa se calcula el correspondiente valor Chi-2 unilateral derecha alfa/2 y unilateral derecha alfa correspondiente a una distribución Chi-Cuadrado con g grados de libertad. Por ejemplo, para $\alpha=0.05$ y g=9 se tiene que:

$$\text{Chi}-2-\text{valor unilateral derecha alfa}/2 = 19.0228 \text{ que verifica}$$

$$2\cdot\text{Prob}\left(\chi^2_g \geq 19.0228\right) = 0.05$$

$$\text{Chi}-2-\text{valor unilateral derecha alfa} = 16.9190 \text{ que verifica}$$

$$\text{Prob}\left(\chi^2_g \geq 16.9190\right) = 0.05$$

Calcular el valor del estadístico Chi-Cuadrado para una probabilidad de 0.05 y 17 grados de libertad.

Resultados de la opción Chi-Cuadrado Inversa.

```
Chi Cuadrado Inversa
===========================================================================

Para alfa = 0.0500 y gl = 17
Chi-2-valor unilateral derecha alfa/2 = 30.1910
Chi-2-valor unilateral derecha alfa = 27.5871
```

Distribuciones → F

Para un valor F de una distribución F con gln grados de libertad del numerador y gld grados de libertad del denominador, se calcula el p-valor unilateral izquierdo y el p-valor unilateral derecho. Por ejemplo, para F=4.3197, gln=6 y gld=9 se tiene que:

$$p-\text{valor unilateral izquierda} = \text{Prob}\left(F_{gln,\ gld} \leq 4.3197\right) = 0.9750$$

$$p-\text{valor unilateral derecha} = \text{Prob}\left(F_{gln,\ gld} \geq 4.3197\right) = 0.0250$$

Calcular la probabilidad de un valor del estadístico F=4.5 para 12 y 2 grados de libertad.

Resultados de la opción F.

```
F
======================================================================

Para F = 4.5000, gln = 12 y gld = 2
p-valor unilateral izquierda = 0.8040
p-valor unilateral derecha = 0.1960
```

Distribuciones → F Inversa

Para un valor alfa se calcula el correspondiente valor F unilateral derecha alfa/2 y unilateral derecha alfa correspondiente a una distribución F con gln grados de libertad del numerador y gld grados de libertad del denominador. Por ejemplo, para α=0.05, gln=6 y gld=9 se tiene que:

$$F-\text{valor unilateral derecha alfa}/2 = 4.3197 \text{ que verifica}$$
$$2 \cdot \text{Prob}\left(F_{gln,\ gld} \geq 4.3197\right) = 0.05$$

$$F-\text{valor unilateral derecha alfa} = 3.3738 \text{ que verifica}$$
$$\text{Prob}\left(F_{gln,\ gld} \geq 3.3738\right) = 0.05$$

Distribuciones → Rango Estudentizado Inversa

Para un valor alfa se calcula el valor Rango Estudentizado unilateral correspondiente a una distribución Rango Estudentizado con gln los grados de libertad del numerador y gld los grados de libertad del denominador. El valor de alfa debe ser para esta opción superior o igual a 0.01 y menor o igual que 0.1.

Por ejemplo para α=0.05, gln=10 y gld=3 se tiene que:

$$\text{Rango Estuden. unilateral} = 3.8774$$

Para α=0.10, gln=10 y gld=3 se tiene que:

$$\text{Rango Estuden. unilateral} = 3.2704$$

Distribuciones → Shapiro Wilk

Para un valor W del estadístico y un tamaño muestral n se calcula el p-valor de la distribución Shapiro-Wilk.

Por ejemplo para W=0.9552 y n=15 se tiene que:

$$p-\text{valor} = 0.3272$$

Cualitativa (a)

Abre un submenú con diferentes técnicas estadísticas para una variable cualitativa dicotómica. Se incluyen las siguientes pruebas: z-proporción y z-proporción para datos agrupados.

Cualitativa (a) → Una proporción

Realiza el intervalo de confianza del parámetro poblacional proporción y un contraste de hipótesis de una proporción mediante la prueba z-proporción.

En el caso de que se esté estudiando una única variable y que ésta sea dicotómica, es recomendable codificar dicha variable con unos y ceros. El valor uno se suele reservar para el código con el que se quiere designar la ocurrencia del suceso de interés, por ejemplo Curación, mientras que el valor de cero se reserva para el suceso complementario, por ejemplo No Curación. Utilizando esta codificación, se expresa una proporción mediante la fórmula:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

El intervalo de confianza de una proporción se calcula como

$$p \in \left[\hat{p} \pm z_{1-\alpha/2} EE(\hat{p})\right]$$

donde

$$EE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

es el error estándar de una proporción y $z_{1-\alpha/2}$ representa el valor de la abscisa en una curva de una distribución normal que deja a la izquierda de su valor un área de 1 - $\alpha/2$ y a la derecha un área de $\alpha/2$. Este valor se encuentra tabulado, de forma que para $\alpha = 0.05$ se tiene que $z_{1-\alpha/2} = 1.96$.

Los intervalos de confianza calculados del modo descrito anteriormente pueden dar lugar a límites sin sentido fuera del intervalo [0 ; 1] en el caso de que se esté trabajando con proporciones extremas cercanas a uno o a cero. Por este motivo, existen métodos más precisos y al mismo tiempo más complejos para calcular los intervalos de confianza para una proporción, como por ejemplo el método exacto de Clopper-Pearson.

En el caso de que se esté interesado en dar afirmaciones acerca de p en términos de, por ejemplo, que la proporción en la población sea un determinado valor p_0, se tiene que plantear el problema en términos de contraste de hipótesis, donde:

$$H_0: p = p_0$$

$$H_1: p \neq p_0$$

siendo H_0 la hipótesis nula y H_1 la hipótesis alternativa.

Para llevar a cabo este contraste se construye el estadístico de contraste experimental z dado por

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

que bajo la hipótesis nula sigue una distribución normal.

Para contrastes unilaterales del tipo:

$$H_0: p = p_0$$

$$H_1: p > p_0$$

se calcula z como antes, pero a la hora de calcular el p-valor, sólo se considera el p-valor como el área bajo la curva normal a la derecha de z (sin considerar el valor absoluto).

En el caso de proporciones extremas conviene usar el método exacto basado en la distribución binomial. En cualquier caso si $n<=50$ se utiliza el método exacto y en caso contrario el asintótico.

Manejo del programa

Variables: Identificar la variable dicotómica a analizar.

Frecuencias: Los principales estadísticos descriptivos para variables cualitativas son: las frecuencias absolutas y las frecuencias relativas. Para cada categoría de la variable cualitativa se muestra el número de individuos que pertenecen a ella (frecuencias absolutas), así como el porcentaje respecto al total de individuos (frecuencias relativas).

Barras: Gráfico de barras para una variable cualitativa. Ver menú Gráficos.

z-Proporción: Se presenta el intervalo de confianza del parámetro proporción poblacional y los resultados del contraste de hipótesis de una proporción mediante la prueba z-proporción.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0.5, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Cualitativa (a) → Una proporción. Datos Agrupados

A partir de los datos resumidos de tamaño muestral y propoción muestral, se puede realizar, sin el fichero de datos, la opción Análisis / Cualitativa (a) / Una proporción. Los fundamentos teóricos y la formulación son idénticos a los presentados en las opciones anteriores respectivas con datos a partir de fichero.

Manejo del programa

Los datos necesarios en el programa son:

- Etiqueta: Nombre de la variable.
- Tamaño Muestral: Valor de n.
- Proporción Muestral: Proporción de la variable en la muestra p.

Una vez introducidos todos los datos, se accede a la pestaña z-Proporción.

z-Proporción: Se presenta el intervalo de confianza del parámetro proporción poblacional y los resultados del contraste de hipótesis de una proporción mediante la prueba z-propoción.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0.5, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Se realiza un contraste de hipótesis sobre la variable Fumador. Los datos necesarios son: Etiqueta=Fumador; Tamaño Muestral= 40; Proporción Muestral= 0.6.

Resultados de la pestaña z-Proporción de la opción Una Proporción.

```
Estimación y Contraste de Una Proporción Poblacional para Fumador
======================================================================

Tamaño Muestral : 40
Proporción:       0.6000

Estimación
----------
I.C.  al 95.00% para la proporción: [0.4333, 0.7514]

Contraste z-Proporción
----------------------
Hipótesis Nula :         proporción = 0.5000
Hipótesis Alternativa :  no igual
p-valor (exacto):        0.2682
```

Cuantitativa (y)

Abre un submenú con diferentes técnicas estadísticas para una variable cuantitativa. Se incluyen las siguientes pruebas: bondad de ajuste, t-Student, Chi-Cuadrado para una desviación típica, t-Student y Chi-Cuadrado para una desviación típica para datos agrupados, rangos signados y signos.

Cuantitativa (y) → Ajuste

Esta opción realiza una prueba de ajuste a una distribución teórica. Se analiza mediante pruebas de bondad de ajuste si se puede asumir que una variable sigue o no la distribución Normal o Uniforme. Las pruebas que se contemplan para el ajuste a una Normal son la prueba de bondad de ajuste Chi-Cuadrado, la prueba de bondad de ajuste de Kolmogorov con corrección de Lilliefors y la prueba de bondad de ajuste de Shapiro-Wilk. Las pruebas que se contemplan para el ajuste a una Uniforme son la prueba de bondad de ajuste Chi-Cuadrado y la prueba de bondad de ajuste de Kolmogorov.

Prueba de bondad de ajuste Chi-Cuadrado

Se divide en tramos el rango de la variable y se analiza el número de valores observados en los distintos tramos y los valores esperados según la distribución teórica correspondiente. El estadístico de contraste se evalúa como la suma en los distintos tramos considerados del cociente entre las diferencias al cuadrado de las frecuencias observadas y esperadas entre las frecuencias esperadas. Este estadístico de contraste sigue una distribución Chi-Cuadrado con grados de libertad igual al número de tramos considerados menos uno y menos el número de parámetros estimados. Este contraste sólo se realiza si n mayor o igual que 20.

Prueba de bondad de ajuste Kolmogorov

Se calcula el estadístico D como la máxima distancia observada entre la función de distribución teórica y la empírica. Para calcular el p-valor, asociado a este estadístico D, es necesario realizar los siguientes cálculos adicionales:

$$Z = \sqrt{n}D$$

$$Q = \mathrm{Exp}\,(-1.233701 / Z^2)$$

$$Q2 = \mathrm{Exp}\,(-2 * Z^2)$$

$$\mathrm{si}\ 0 <= Z < 0.27 \Rightarrow p\text{ - valor} = 1$$

$$\mathrm{si}\ 0.27 <= Z < 1 \Rightarrow p\text{ - valor} = 1 - (2.506628/Z) * (Q + Q^9 + Q^{25})$$

$$\mathrm{si}\ 1 <= Z < 3.1 \Rightarrow p\text{ - valor} = 2 * (Q2 - Q2^4 + Q2^9 - Q2^{16})$$

$$\mathrm{si}\ Z >= 3.1 \Rightarrow p\text{ - valor} = 0$$

Prueba de bondad de ajuste de Kolmogorov con corrección de Lilliefors

En el caso de realizar la prueba de Kolmogorov de ajuste a la normal, es recomendable utilizar la corrección de Lilliefors, con lo que el p-valor de Lilliefors p_L se calcula de la forma siguiente:

Si n <= 100 $\Rightarrow$

$$A = -7.01256 * (n + 2.78019)$$

$$B = 2.99587 * \sqrt{n + 2.78019}$$

$$C = 2.1804661 + 0.974598/\sqrt{n} + 1.67997/n$$

Si n > 100 $\Rightarrow$

$$A = -7.90289126054 * n^{0.98}$$

$$B = 3.180370175721 * n^{0.49}$$

$$C = 2.2947256$$

Y se calcula DC de forma que

$$DC = \frac{-B - \sqrt{B * B - 4 * A * C}}{2 * A}$$

$$\mathrm{si}\ D = DC \Rightarrow p_L = 0.1$$

$$\mathrm{si}\ D > DC \Rightarrow p_L = \mathrm{Exp}\,(A * D^2 + B * D + C - 2.3025851)$$

$$\mathrm{si}\ D < DC \Rightarrow p_L = "> 0.1"$$

Prueba de bondad de ajuste de Shapiro-Wilk

Shapiro y Wilk (1965) introducen la prueba W de normalidad. Posteriormente, en 1982, Royston implementa un algoritmo con el nombre de AS 181 para $7 \leq n \leq 2000$ para llevar a cabo dicha prueba. En 1992 el propio Royston descubre que el algoritmo AS 181 es incorrecto para $n > 50$ y en 1995 propone un nuevo algoritmo con el nombre de AS R94 válido para $3 \leq n \leq 5000$. Hasta el año 2000 los principales programas estadísticos comerciales no corrigieron el error y siguieron usando el algoritmo AS 181. En G-Stat está implementada la versión corregida AS R94. Se puede decir, por tanto, que la técnica más moderna de G-Stat es el algoritmo AS R94 para el cálculo de la prueba W de Shapiro-Wilk.

Manejo del programa

Variables: Se identifica la variable cuantitativa para ver su posible distribución.

Estadísticos: Estadísticos de la variable seleccionada. Se detallan en Cuantitativa (y).

Histograma: Genera un histograma. Más información en el apartado Menú Gráficos.

Contraste: Resultados de la prueba de bondad de ajuste de una variable a una distribución teórica.

Opciones:

- Distribución: Normal o uniforme.

Se desea comprobar si la variable FC2FC1 sigue una distribución Normal.

Resultados del contraste de normalidad de la variable FC2FC1.

```
Contrastes de Hipótesis de Bondad de Ajuste para FC2FC1
===========================================================================
Número de Casos:          40

Distribución Teórica:     Normal
Media:                    61.7500
Desviación Típica:        10.9772
```

Límite Inferior	Límite Superior	Frec. Observadas	Frec. Esperadas	Chi-Cuadrado
inferior	51.1304	7	6.6667	0.0167
51.1304	57.0218	6	6.6667	0.0667
57.0218	61.7500	7	6.6667	0.0167
61.7500	66.4782	6	6.6667	0.0667
66.4782	72.3696	7	6.6667	0.0167
72.3696	superior	7	6.6667	0.0167

```
Chi Cuadrado = 0.2000 con 3.0 G.L. p-valor = 0.9776

D+ de Kolmogorov:          0.0633
D- de Kolmogorov:          -0.0738
DN:                        0.0738
p-valor:                   0.9812

p-valor Lilliefors corregido:  >0.1

W Shapiro-Wilk:                0.9753
p-valor Shapiro-Wilk:          0.5212
```

La prueba recomendada es la de Shapiro-Wilk.

Cuantitativa (y) → t-Student

Realiza el intervalo de confianza del parámetro poblacional media y un contraste de hipótesis mediante la prueba t-Student para una muestra.

El intervalo de confianza para una media se calcula como

$$\mu \in \left[\bar{x} \pm t_{1-\alpha/2;gl} EE(\bar{x})\right]$$

donde

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)}$$

$$EE\left(\bar{x}\right) = \frac{s}{\sqrt{n}}$$

siendo $EE\left(\bar{x}\right)$ el error estándar de la media y $t_{1-\alpha/2;\ gl}$ representa el valor de la abscisa en una curva de una distribución t-Student con gl grados de libertad dados por (n - 1), que deja a la izquierda de su valor un área de 1 - α/2 y a la derecha un área de α/2. Este valor se encuentra tabulado, de forma que para α = 0.05 y gl = 9 se tiene que $t_{1-\alpha/2;gl}$ = 2.26.

En el caso de que se esté interesado en dar afirmaciones acerca de μ en términos de, por ejemplo, que la media en la población sea un determinado valor μ_0, se tiene que plantear el problema en términos de contraste de hipótesis, donde:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

siendo H_0 la hipótesis nula y H_1 la hipótesis alternativa.

Para llevar a cabo este contraste, se construye el estadístico de contraste experimental t dado por

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

que recibe el nombre de prueba t-Student para una media y que bajo la hipótesis nula sigue una distribución t-Student con (n - 1) grados de libertad. Para contrastes bilaterales se calcula el p-valor como el área bajo la curva t-Student con (n - 1) grados de libertad a la derecha del valor absoluto de t más el área a la izquierda de menos el valor absoluto de t.

Para contrastes unilaterales del tipo:

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

se calcula t como antes, pero a la hora de calcular el p-valor sólo se considera el área bajo la curva de la distribución normal a la derecha de t (sin considerar el valor absoluto).

Manejo del programa

Variables: Se identifica la variable cuantitativa a analizar.

Estadísticos: Estadísticos de la variable seleccionada. Se detallan en Cuantitativa (y).

Cajas: Gráfico de Cajas para una variable cuantitativa. Ver menú Gráficos

t-Student: Se presenta el intervalo de confianza del parámetro poblacional media y los resultados del contraste de hipótesis de una media mediante la prueba t-Student para una muestra.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0.0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Contrastar si la media de la variable FC1FC2 es superior a 60.

Resultados de la prueba t-Student para una variable.

```
Estimación y Contraste de Una Media Poblacional para FC2FC1
======================================================================

Tamaño Muestral:   40
Media:             61.7500

Estimación
----------
I.C. inferior al 95.0000% para la media:  61.7500 - 2.9244 [58.8256]

t-Student
---------
Hipótesis Nula:                 media = 60.0000
Hipótesis Alternativa:          mayor que
Estadístico de contraste t:     1.0083
p-valor:                        0.1598
```

Cuantitativa (y) → Chi-2 para una Desviación Típica

Realiza el intervalo de confianza del parámetro poblacional desviación típica y un contraste de hipótesis de una desviación típica mediante el estadístico de Chi-Cuadrado.

El intervalo de confianza de una desviación típica se calcula como

$$\sigma \in \left(\sqrt{\frac{\sum \left(x_i - \overline{x}\right)^2}{\chi^2_{gl,1-\alpha/2}}}, \sqrt{\frac{\sum \left(x_i - \overline{x}\right)^2}{\chi^2_{gl,\alpha/2}}} \right)$$

donde $\chi^2_{gl,1-\alpha/2}$ representa el valor de la abscisa en una curva de una distribución Chi-Cuadrado con grados de libertad gl=n-1 que deja a la izquierda de su valor un área de 1 - $\alpha/2$ y $\chi^2_{gl,\alpha/2}$ representa el valor de la abscisa en una curva de una distribución Chi-Cuadrado con grados de libertad gl que deja a la izquierda de su valor un área de $\alpha/2$, de forma que para α = 0.05 y n=10, se tiene que $\chi^2_{gl,1-\alpha/2} = 19.0228$ y $\chi^2_{gl,\alpha/2} = 2.7004$.

En el caso de que se esté interesado en dar afirmaciones acerca de σ en términos de, por ejemplo, que la desviación típica en la población sea un determinado valor σ_0, se tiene que plantear el problema en términos de contraste de hipótesis, donde:

$$H_0: \sigma = \sigma_0$$

$$H_1: \sigma \neq \sigma_0$$

siendo H_0 la hipótesis nula y H_1 la hipótesis alternativa.

Para llevar a cabo este contraste se construye el estadístico de contraste experimental χ^2 dado por

$$\chi^2 = \frac{\left(n-1\right)s^2}{\sigma_0^2} = \frac{\sum \left(x_i - \overline{x}\right)^2}{\sigma_0^2}$$

que bajo la hipótesis nula sigue una distribución Chi-Cuadrado con grados de libertad gl = n - 1.

Manejo del programa

Variables: Se identifica la variable cuantitativa a analizar.

Estadísticos: Estadísticos de la variable seleccionada. Se detallan en Cuantitativa (y).

Cajas: Gráfico de Cajas para una variable cuantitativa. Ver menú Gráficos

Chi-Cuadrado: Se presenta el intervalo de confianza del parámetro poblacional desviación típica y los resultados del contraste de hipótesis de una desviación típica mediante el estadístico de Chi-Cuadrado.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 1.0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Se quiere contrastar si la desviación típica de la variable FC2FC1 es superior a un valor teórico de 8.

Resultados de la prueba Chi-2 para una desviación típica.

```
Estimación y Contraste de Una Desviación Típica Poblacional para FC2FC1
=======================================================================

Tamaño Muestral:           40
Media:                     61.7500
Desviación Típica:         10.9772
Grados de Libertad:        39

Estimación
----------
I.C. inferior al 95.00% para la desviación típica: [9.2798]

Contraste Chi-Cuadrado
----------------------
```

```
Hipótesis Nula:                                  sigma = 8.0000
Hipótesis Alternativa:                           mayor que
Estadístico de contraste chi-cuadrado:           73.4297
p-valor:                                         0.0007
```

Cuantitativa (y) → t-Student y Chi-2 para dt. Datos Agrupados

A partir de los datos resumidos de tamaño muestral, media y desviación típica muestral, se puede realizar, sin el fichero de datos, las opciones Análisis / Cuantitativa (y) / t-Student y Análisis/ Cuantitativa (y) / Chi-2 para una Desviación Típica. Los fundamentos teóricos y la formulación son idénticos a los presentados en las opciones anteriores respectivas con datos a partir de fichero.

Se tiene en cuenta que aunque no se tengan los valores individuales x_i, se verifica que

$$\sum \left(x_i - \overline{x}\right)^2 = \left(n-1\right) s^2$$

Manejo del programa

Los datos necesarios son:

- Etiqueta: Nombre variable.
- Tamaño muestral: el valor de n.
- Media muestral
- Desviación típica muestral.

Se accede a las pestañas t-Student y Chi-Cuadrado. Es necesario haber cumplimentado todos los datos.

t-Student: Se presenta el intervalo de confianza del parámetro poblacional media y los resultados del contraste de hipótesis de una media mediante la prueba t-Student para una muestra.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0.0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Chi-Cuadrado: Se presenta el intervalo de confianza del parámetro poblacional desviación típica y los resultados del contraste de hipótesis de una desviación típica mediante el estadístico de Chi-Cuadrado.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 1.0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Cuantitativa (y) → Rangos Signados

Realiza un contraste de hipótesis de la mediana de una distribución mediante la prueba de los Rangos signados. Es una prueba no paramétrica.

La prueba de los Rangos signados para una muestra calcula la variable Rta2 como en el caso de la prueba de Signos, a continuación se crea la variable Orden del valor absoluto de Rta2, que es el orden que ocupa la variable Valor absoluto de Rta2, después se calcula los rangos de Orden del valor absoluto de Rta2, asignando el rango medio en caso de empates poniendo el signo que tuviera la variable Rta2 (de ahí el nombre de Rangos signados). Como antes:

n_+ = número de signos (+) de Rta2

n_- = número de signos (-) de Rta2

$$n^* = n_+ + n_-$$

A partir de estos rangos se consideran

$$T_+ = \sum_{grupo+} rangos,$$

$$T_- = \sum_{grupo-} rangos,$$

que verifican

$$E[T_+] = E[T_-] = \frac{1}{4} n^* (n^* + 1),$$

$$V[T_+] = V[T_-] = \sqrt{\frac{1}{24} n^*(n^* + 1)(2n^* + 1) - \sum_{empates} \frac{1}{48}(n^o\ empates^3 - n^o\ empates)}$$

La forma de realizar los contrastes bilaterales y unilaterales se recoge en la siguiente tabla:

Hipótesis nula	Hipótesis alternativa	T
med = med_0	med ≠ med_0	mín {T+ T-}
med = med_0	med < med_0	T+
med = med_0	med > med_0	T-

En todas las situaciones el estadístico de contraste es

$$z = \frac{T - E[T]}{\sqrt{V[T]}}$$

que sigue una distribución normal N(0,1).

Manejo del programa

Variables: Se identifica la variable cuantitativa a analizar.

Estadísticos: Estadísticos de la variable seleccionada. Se detallan en Cuantitativa (y).

Cajas: Gráfico de Cajas para una variable cuantitativa. Ver menú Gráficos

Rangos Signados: Se presentan los resultados del contraste de hipótesis de la mediana de una distribución mediante la prueba de Rangos signados.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0.0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

Contrastar, mediante la prueba de Rangos signados, si la mediana de la variable FC2 es superior a 130.

Resultado de la prueba de los rangos signados para una variable.

```
Prueba de Rangos Signados de la mediana para FC2
======================================================================
Tamaño Muestral:   40
Mediana:           137.0000

Rangos Signados
---------------

Hipótesis Nula:               mediana = 130.0000
Hipótesis Alternativa:        no igual

Rango medio de valores por debajo del valor a contrastar de la mediana:
22.9464
Rango medio de valores por encima del valor a contrastar de la mediana:
14.7917

Estadístico de contraste:      -3.1287 (con corrección por empates)
p-valor:                       0.0018
```

Cuantitativa (y) → Signos

Realiza un contraste de hipótesis de la mediana de una distribución mediante la prueba de los signos. Es una prueba no paramétrica.

La prueba de los Signos para una muestra calcula, en primer lugar, una nueva variable, a la que se denota por Rta2, que es igual a la variable original, que se denota por Rta, menos el valor que se está contrastando med_0. Para esta nueva variable se calculan:

$$n_+ = \text{número de signos } (+) \text{ de Rta2}$$

$$n_- = \text{número de signos } (-) \text{ de Rta2}$$

$$n^* = n_+ + n_-$$

con lo que n_+ sería el número de valores de Rta por encima del valor de la mediana que se esté contrastando y n_- sería el número de valores de Rta por debajo. Observar que n^* es menor que el tamaño original, ya que no se cuentan los empates producidos por aquellos valores de individuos con valor en la variable original igual a med_0.

La forma de realizar los contrastes bilaterales y unilaterales se recoge en la siguiente tabla:

Hipótesis nula	Hipótesis alternativa	k
$med = med_0$	$med \neq med_0$	mín $\{n_+, n_-\}$
$med = med_0$	$med < med_0$	n_+
$med = med_0$	$med > med_0$	n_-

En todas las situaciones se considera

$$E[k] = \frac{1}{2} n^*$$

$$V[k] = \frac{1}{2}\frac{1}{2} n^*$$

y el estadístico de contraste es

$$z = \frac{k - E[k]}{\sqrt{V[k]}}$$

que sigue una distribución normal N(0,1).

Manejo del programa

Variables: Se identifica la variable cuantitativa a analizar.

Estadísticos: Estadísticos de la variable seleccionada. Se detallan en Cuantitativa (y).

Cajas: Gráfico de Cajas para una variable cuantitativa. Ver menú Gráficos

Signos: Se presentan los resultados del contraste de hipótesis de la mediana de una distribución mediante la prueba de signos.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0.0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

Contrastar, mediante la prueba de los signos, si la mediana de la variable FC2 es superior a 130.

Resultados de la prueba de los signos para una variable.

```
Prueba de Signos de la mediana para FC2
========================================================================

Tamaño Muestral:        40
Mediana:                137.0000

Signos
------
Hipótesis Nula:               mediana = 130.0000
Hipótesis Alternativa:        mayor que

Número de valores por debajo del valor a contrastar de la mediana:   12
Número de valores por encima del valor a contrastar de la mediana:   28

Estadístico de contraste:         -2.5298
p-valor:                          0.0057
```

Tablas (a|b)

Contiene un submenú con diferentes técnicas estadísticas para dos variables cualitativas. En algunos casos se supone que existe un modelo donde una variable explicativa cualitativa (a) explica una variable respuesta cualitativa (b); en otros casos no se supone ningún modelo y estamos ante técnicas de asociación. Se incluyen las siguientes pruebas: Chi-Cuadrado, Chi-Cuadrado para datos agrupados, z-proporciones para datos agrupados, Chi-Cuadrado de tendencia lineal, Chi-Cuadrado de tendencia lineal para datos agrupados, Fisher, Fisher para datos agrupados, McNemar y McNemar para datos agrupados.

Tablas (a|b) → Chi-Cuadrado

Realiza la prueba Chi-Cuadrado que contrasta la asociación entre dos variables cualitativas que pueden ser dicotómicas o nominales. En el caso de que una de las variables haga el papel de variable respuesta y sea dicotómica, estamos hablando de la comparación de varias proporciones. Se necesita que no más del 20% de las celdas tengan valores esperados menores de cinco.

Para calcular el estadístico de contraste Chi-Cuadrado, se construye en primer lugar la tabla de contingencia de dimensiones r (número de filas) por c (número de columnas) con las frecuencias absolutas observadas n_{ij}, que son el resultado de contar el número de individuos para cada par de posibilidades de los distintos niveles i de la variable en filas y j de la variable en columnas.

A continuación se calcula la tabla de contingencia de frecuencias absolutas esperadas mediante la expresión:

$$e_{ij} = \frac{r_i c_j}{n}$$

donde r_i indica el total de individuos de la fila i , c_j el total de individuos de la columna j y n el total de individuos.

El estadístico de contraste es

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

que sigue bajo la hipótesis nula una distribución χ^2, con (r-1)x(c-1) grados de libertad.

Manejo del programa

Variables: Se identifica la *Variable Fila* y la *Variable Columna*.

Tablas: Se muestra la tabla de contingencia de dos variables cualitativas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Chi-Cuadrado: Se presentan los resultados de la prueba Chi-Cuadrado que contrasta la asociación entre dos variables cualitativas. Asimismo, se muestra, para ver la validez de los resultados, el número y proporción de celdas con frecuencias absolutas esperadas menor de 5 y menor de 1.

Realizar la prueba Chi-Cuadrado para establecer si la variable Sexo está relacionada con la variable Fumador.

Tabla de frecuencias de la opción Chi-Cuadrado.

```
Tabla de Frecuencias de FUMADOR (filas) por SEXO (columnas)
======================================================================
Número de Casos: 40

 SEXO          |    1         |    2         |    Total
 FUMADOR       |              |              |    Fila
------------------------------------------------------
               |    10        |    6         |    16
 1             |    62.50     |    37.50     |    40.00
------------------------------------------------------
               |    12        |    12        |    24
 2             |    50.00     |    50.00     |    60.00
------------------------------------------------------
 Total         |    22        |    18        |    40
  Columna      |    55.00     |    45.00     |    100.00

Los porcentajes de cada celda se refieren al total de cada fila
```

Resultados de la prueba Chi-Cuadrado.

```
Chi-Cuadrado de FUMADOR (filas) por SEXO (columnas)
======================================================================

Tamaño Muestral:                       40
Estadístico de contraste Chi-Cuadrado: 0.6061
G.L.:                                  1
p-valor:                               0.4363
Nº de celdas con frecuencias absolutas esperadas < 5: 0 de 4, un 0.0000%
Nº de celdas con frecuencias absolutas esperadas < 1: 0 de 4, un 0.0000%
```

Tablas (a|b) → Chi-Cuadrado. Datos Agrupados

A partir de datos agrupados se realiza la prueba de Chi-Cuadrado para dos variables cualitativas. Los datos agrupados se introducen directamente en una cuadrícula en forma de tabla de r filas y c columnas. Los fundamentos teóricos y la formulación son idénticos a los presentados en la opción de análisis anterior con datos a partir de un fichero.

Manejo del programa

Datos Agrupados: En la pestaña de datos agrupados hay una cuadrícula que permite la entrada directa del número de casos dentro de cada posible combinación de categorías o niveles. Se puede definir el número de niveles de las dos variables mediante el número de filas y columnas. El botón "Crear Tabla" prepara la estructura de la tabla ajustada al número de filas y columnas definido. Por defecto aparecen en la tabla unos valores que deben ser sustituidos por los datos del usuario.

No dejar en la tabla filas o columnas con todos los valores faltantes o con todos los valores iguales a cero. Las celdas de la tabla no admiten valores faltantes, negativos, decimales o alfanuméricos. Este programa no permite tablas menores de 2x2.

Tablas: Se muestra la tabla de contingencia de dos variables cualitativas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías. Estos valores son los que se han introducido en la cuadrícula de entrada de datos. Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Chi-Cuadrado: Se presentan los resultados de la prueba Chi-Cuadrado que contrasta la asociación entre dos variables cualitativas. Asimismo, se muestra, para ver la validez de los resultados, el número y proporción de celdas con frecuencias absolutas esperadas menor de 5 y menor de 1.

Se desea realizar la prueba Chi-Cuadrado sobre la siguiente tabla de contingencia:

	Var a_1	Var a_2	Var a_3
Var b_1	23	45	12
Var b_2	34	66	26

En la pestaña Datos Agrupados crear una tabla de 2 filas y 3 columnas, introducir los datos, hacer clic en cualquier otra celda de la tabla agregada para confirmar todos los datos e ir a la pestaña Chi-Cuadrado.

Resultados de la prueba Chi-Cuadrado.

```
Chi-Cuadrado de Filas por Columnas
========================================================================

Tamaño Muestral:                       206
Estadístico de contraste Chi-Cuadrado: 1.0334
G.L.:                                  2
p-valor:                               0.5965
Nº de celdas con frecuencias absolutas esperadas < 5: 0 de 6, un 0.0000%
Nº de celdas con frecuencias absolutas esperadas < 1: 0 de 6, un 0.0000%
```

Tablas (a|b) → Dos Proporciones. Datos Agrupados

A partir de datos agrupados se realiza el intervalo de confianza del parámetro diferencia poblacional de dos proporciones y el contraste de hipótesis de dos proporciones mediante la prueba z-Proporciones.

Si se está en el caso de que se tengan dos variables dicotómicas donde se pueda considerar una variable explicativa, por ejemplo, Tratamiento con dos niveles dados por Tratamiento A y Tratamiento B y otra variable respuesta, por ejemplo, Estado de salud con dos niveles dados por Enfermo y Sano, se podría ver si hay influencia de la variable explicativa Tratamiento en la variable respuesta Estado de salud. Para ello se estudiaría la diferencia entre la proporción de enfermos para el Tratamiento A y la proporción de enfermos para el Tratamiento B.

La forma de realizar un intervalo de confianza para el verdadero valor del parámetro diferencia de dos proporciones $p_1 - p_2$ es

$$p \in \left[\left(\hat{p}_1 - \hat{p}_2\right) \pm z_{1-\alpha/2} EE(\hat{p}_1 - \hat{p}_2)\right]$$

donde $\hat{p}_1$ representa la proporción muestral estimada para p_1 y $\hat{p}_2$ representa la proporción muestral estimada para p_2 y $EE(\hat{p}_1 - \hat{p}_2)$ el error estándar estimado de la diferencia de dos proporciones (la raíz cuadrada de la varianza de la distribución muestral de la diferencia de dos proporciones) dado por

$$EE(\hat{p}_1 - \hat{p}_2) = \sqrt{EE(\hat{p}_1)^2 + EE(\hat{p}_2)^2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

que recoge la variabilidad muestral.

En el caso de que se quiera realizar contraste de hipótesis para dos proporciones del tipo:

$$H_0: p_1 - p_2 = p_0$$

$$H_1: p_1 - p_2 \neq p_0$$

siendo H_0 la hipótesis nula y H_1 la hipótesis alternativa, es necesario construir el estadístico de contraste experimental z dado por

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{EE(\hat{p}_1 - \hat{p}_2)}$$

Bajo la hipótesis nula $p_0 = 0$, las dos proporciones son iguales y se tiene que

$$EE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} = \sqrt{\hat{p}(1-\hat{p})\left[(1/n_1) + (1/n_2)\right]}$$

donde $\hat{p}$ está dada por:

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

y representa una proporción común de individuos que tienen presente la característica de interés sin distinguir por grupos.

En el caso de que $p_0 \neq 0$, se tiene que

$$EE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

proporciona el error estándar de la diferencia de dos proporciones.

En el caso de proporciones extremas conviene usar el método exacto basado en la distribución binomial.

Manejo del programa

Datos Agrupados: Los datos necesarios en el programa son:

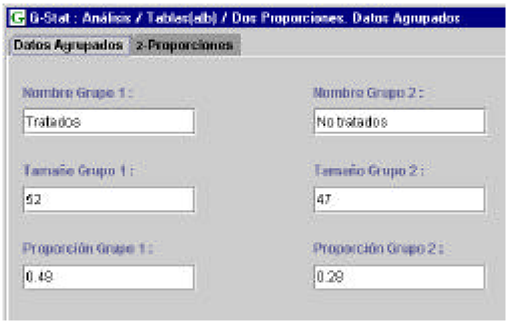
- Nombre Grupo 1: Etiqueta Categoría1.
- Nombre Grupo 2: Etiqueta Categoría2.
- Tamaño Grupo1: Valor de n1.
- Tamaño Grupo2: Valor de n2.
- Proporción Grupo1: Valor de p1.
- Proporción Grupo2: Valor de p2.

z-Proporciones: Resultados de la estimación y contraste de la diferencia de dos proporciones.

Opciones:

- Permite cambiar los valores que por defecto aparecen, Hipótesis nula: 0.0, Hipótesis alternativa: (no igual, menor que, mayor que) y alfa: 5.0%.

Comparar una proporción del 48% con un tamaño de muestra de 52, frente a una del 28% con un tamaño de muestra de 47.



Entrada de datos para la comparación de dos proporciones.

Resultados de la opción comparación de dos proporciones.

```
Estimación y Contraste de la Diferencia Poblacional de Dos Proporciones
======================================================================

Tamaños Muestrales:                                 52 y 47
Proporciones:                                       0.4800 y 0.2800
Error estándar de las proporciones:                 0.0693 y 0.0655
Diferencia de proporciones:                         0.2000
Error estándar de la diferencia de proporciones:    0.0953

Estimación
----------
I.C.  al 95.00% para la diferencia de dos proporciones:
0.2000 +/- 0.1869   [0.0131, 0.3869]

Contraste z-Proporciones
------------------------
Hipótesis Nula:             proporción1-proporción2 = 0.0000
Hipótesis Alternativa:      no igual
Estadístico de contraste z: 2.0421
p-valor:                    0.0411
```

Tablas (a|b) → Chi-Cuadrado de Tendencia Lineal (y|b)

Realiza la prueba Chi-Cuadrado de tendencia lineal de proporciones de una variable respuesta dicotómica (b) a partir de los subgrupos formados por una variable explicativa discreta cuantitativa (y).

El estadístico de contraste Chi-Cuadrado de tendencia lineal es un componente del valor del estadístico Chi-Cuadrado para tablas de frecuencias.

La notación que se sigue es la de una matriz con 2 filas y k columnas donde:

	Var. en columnas=var. explicativa X				
Var. en filas=var. respuesta	x_1	x_2	...	x_k	Total
1="Presencia"	y_1	y_2	...	y_k	y
0="Ausencia"	$(n_1\text{-}y_1)$	$(n_1\text{-}y_1)$	...	$(n_k\text{-}y_k)$	n-y
Total	n_1	n_2	...	n_c	n

El estadístico de contraste Chi-Cuadrado de tendencia lineal viene dado por

$$\chi_t^2 = \frac{n\left(n\sum_{i=1}^{k} y_i x_i - y\sum_{i=1}^{k} n_i x_i\right)^2}{y(n-y)\left[n\sum_{i=1}^{k} n_i x_i^2 - \left(\sum_{i=1}^{k} n_i x_i\right)^2\right]}$$

que sigue una Chi-Cuadrado con un grado de libertad.

Se observa que las "etiquetas" de la variable explicativa intervienen en el cálculo del estadístico de contraste, con lo que éste variará si se cambia la codificación de la variable respuesta.

Manejo del programa

Variables: Se identifica la *Variable Fila* y la *Variable Columna*.

Tablas: Se muestra la tabla de contingencia de dos variables cualitativas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Chi-Cuadrado de Tendencia Lineal: Contrasta la linealidad (creciente o decreciente) de la respuesta en función de la variable formadora de grupos.

Realizar la prueba Chi-Cuadrado de tendencia lineal para establecer si la variable Sexo está relacionada linealmente con la variable Status.

Tabla de frecuencias

Tabla de Frecuencias de SEXO (filas) por STATUS (columnas)

Número de Casos: 40

STATUS SEXO	1	2	3	Total Fila
2	10 58.82	7 63.64	1 8.33	18 45.00
1	7 41.18	4 36.36	11 91.67	22 55.00
Total Columna	17 42.50	11 27.50	12 30.00	40 100.00

Los porcentajes de cada celda se refieren al total de cada columna

Resultados de la prueba Chi-Cuadrado de tendencia lineal.

Chi-Cuadrado de tendencias de SEXO (filas) por STATUS (columnas)

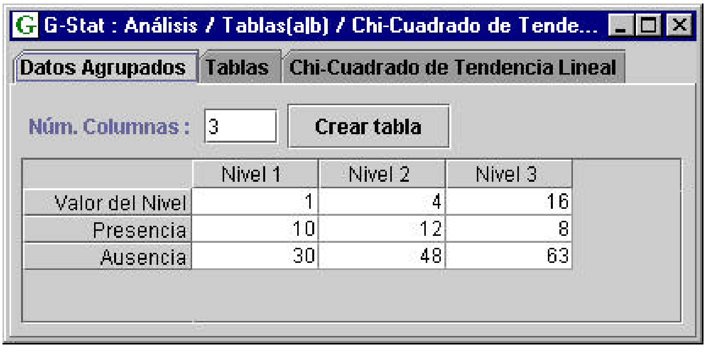
Tamaño Muestral:	40
Estadístico de contraste Chi-Cuadrado de tendencias:	6.4878
G.L.:	1
p-valor:	0.0109

Tablas (a|b) → Chi-Cuadrado de Tendencia Lineal. Datos Agrupados (y|b)

A partir de datos agrupados se realiza la prueba Chi-Cuadrado de tendencia lineal de proporciones. Los datos agrupados se introducen directamente en una cuadrícula en forma de tabla de (2+1) filas y c columnas. Los fundamentos teóricos y la formulación son idénticos a los presentados en la opción de análisis anterior con datos a partir de un fichero.

Manejo del programa

Datos Agrupados: La disposición de la tabla para la entrada de datos es la siguiente:



En la pestaña de datos agrupados hay una cuadrícula que permite la entrada directa del número de casos dentro de cada posible combinación de categorías o niveles. Sólo se puede definir el número de categorías de la variable explicativa en columnas, ya que la variable respuesta siempre es dicotómica con valores de "presencia" y "ausencia". El botón "Crear Tabla" prepara la estructura de la tabla ajustada al número de columnas definido. Por defecto aparecen en la tabla unos valores que deben ser sustituidos por los datos del usuario. Se puede asociar un valor numérico a las categorías o niveles de la variable explicativa ordinal que va en columnas y que se introducen en la primera fila de la tabla. En las dos filas siguientes de la tabla se especifican el número de casos para cada nivel de la variable explicativa según la presencia o ausencia de la variable respuesta.

No dejar en la tabla filas o columnas con todos los valores faltantes o con todos los valores iguales a cero. Las celdas de la tabla no admiten valores faltantes, negativos, decimales o alfanuméricos. Este programa no permite tablas menores de 2x2.

Tablas: Se muestra la tabla de contingencia de dos variables cualitativas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Chi-Cuadrado de Tendencia Lineal: Contrasta la linealidad (creciente o decreciente) de la respuesta en función de la variable formadora de grupos.

Tablas (a|b) → Fisher

Realiza la prueba de Fisher que contrasta la asociación entre dos variables dicotómicas. En el caso de que una de las variables haga el papel de variable respuesta, estamos hablando de la comparación de dos proporciones. Se emplea para muestras pequeñas donde no se da el supuesto de que no haya más del 20% de las celdas con valores esperados menores de cinco. La prueba de Fisher utiliza cálculos exactos pues no se calcula la significación mediante la aproximación asintótica.

Para calcular el estadístico de contraste, se construye la tabla de contingencia de dimensiones 2x2 con las frecuencias absolutas observadas, con la notación siguiente:

	Var1		
Var2	Cat1	Cat2	Total
Niv1	a	b	r_1
Niv2	c	d	r_2
Total	c_1	c_2	n

A continuación, se construyen todas las tablas de contingencia 2x2 posibles con celdas a', b', c', d', siendo $0 \leq a' \leq \text{mín}\{c_1, r_1\}$, $b' = r_1 - a'$, $c' = c_1 - a'$ y $d' = r_2 - c'$. A partir de dichas tablas se calcula:

$$p_{a'} = \frac{r_1!\, r_2!\, c_1!\, c_2!}{n!\, a'!\, b'!\, c'!\, d'!}$$

donde x! indica el factorial de x, que se calcula como x!=x·(x-1)·(x-2)·…·2·1, por ejemplo, 5!=5·4·3·2·1=120.

El p-valor unilateral izquierdo resultante es

$$\sum_{a' \leq a} p_{a'}$$

el p-valor unilateral derecho es

$$\sum_{a' \geq a} p_{a'}$$

y el p-valor bilateral es

$$\sum_{p_{a'} \leq p_a} p_{a'}$$

Manejo del programa

Variables: Se identifica la *Variable Fila* y la *Variable Columna*. En G-Stat, esta prueba se restringe a variables dicotómicas.

Asegurarse en la pestaña de tablas que el orden de las categorías es el deseado. Si no es así, ordenar los datos para conseguir la disposición deseada.

Tablas: Muestra la tabla de contingencia de dos variables cualitativas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total).

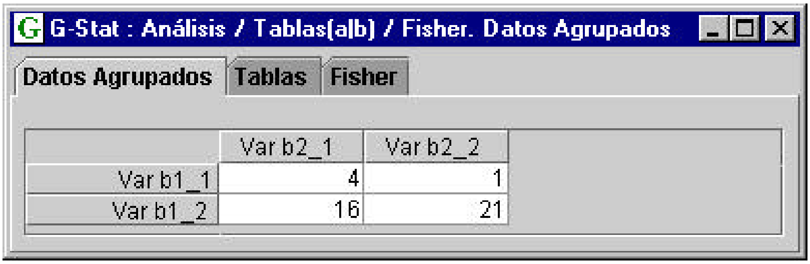
Fisher: Se presentan los resultados de la prueba de Fisher. Presenta los valores de los contrastes unilaterales y el bilateral.

Tablas (a|b) → Fisher. Datos agrupados

Contrasta la asociación entre dos variables dicotómicas mediante la prueba exacta de Fisher. Se emplea para muestras pequeñas donde no se da el supuesto de que no haya más del 20% de las celdas con valores esperados menores de cinco. Los datos se introducen agrupados en frecuencias en la cuadrícula correspondiente. Los fundamentos teóricos y la formulación son idénticos a los presentados en la opción de análisis anterior con datos a partir de un fichero.

Manejo del programa

Datos Agrupados: La disposición de la tabla para la entrada de datos es la siguiente:



En la pestaña de datos agrupados hay una cuadrícula que permite la entrada directa del número de casos dentro de cada posible combinación de categorías o niveles. Las variables deben ser dicotómicas, con lo que la tabla resultante es siempre dos por dos. Por defecto aparecen en la tabla unos valores que deben ser sustituidos por los datos del usuario.

No dejar en la tabla filas o columnas con todos los valores faltantes o con todos los valores iguales a cero. Las celdas de la tabla no admiten valores faltantes, negativos, decimales o alfanuméricos. Este programa no permite tablas menores de 2x2.

Asegurarse en la pestaña de tablas que el orden de las categorías es el deseado. Si no es así, ordenar los datos para conseguir la disposición deseada.

Tablas: Muestra la tabla de contingencia de dos variables cualitativas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Fisher: Se presentan los resultados de la prueba de Fisher. Presenta los valores de los contrastes unilaterales y el bilateral.

Tablas (a|b) → McNemar

Realiza la prueba de McNemar para datos pareados que compara dos proporciones de dos muestras pareadas. Es una prueba no paramétrica. Adicionalmente, también se realiza el cálculo del coeficiente kappa de concordancia y de su significación estadística.

Los datos se expresan en una tabla de contingencia 2x2, donde en cada celda están los pares correspondientes a cada cruce de categoría de las dos variables dicotómicas. Los pares discordantes (fuera de la diagonal principal) son los que influyen en la prueba, la notación es la siguiente:

	Var2		
Var1	Cat1	Cat2	Total
Cat1	a	b	r_1
Cat2	c	d	r_2
Total	c_1	c_2	n

El estadístico de contraste se construye como

$$z = \frac{b - \frac{1}{2}(b+c)}{\frac{1}{2}\sqrt{b+c}} = \frac{\frac{1}{2}(b-c)}{\frac{1}{2}\sqrt{b+c}} = \frac{b-c}{\sqrt{b+c}}$$

que sigue una distribución normal N(0,1). Este programa proporciona un p-valor asintótico para esta opción.

Alternativamente, se puede considerar el estadístico de contraste:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

que sigue una distribución Chi-Cuadrado con un grado de libertad y que tiene asociado el mismo p-valor ya que se cumple que

$$z^2 = \chi^2$$

con lo que ambos procedimientos son equivalentes.

Adicionalmente, el programa muestra el estadístico kappa de concordancia, que viene dado por

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

con

$$p_0 = \frac{a + d}{n}\ ,\ \ p_c = \frac{r_1 c_1 - r_2 c_2}{n^2}$$

El estadístico de contraste se construye como

$$z = \frac{\kappa}{\sqrt{\frac{p_c}{n(1 - p_c)}}}$$

que sigue una distribución normal N(0,1). Este programa proporciona un p-valor asintótico para esta opción.

Manejo del programa

Variables: Se identifica la *1º Variable Pareada* y la *2ª Variable Pareada,* ambas dicotómicas.

Se asume, para el análisis de los datos, que la diagonal de concordancia viene dada por la casilla superior izquierda y la casilla inferior derecha. Si no es así, ordenar los datos para conseguir dicha disposición. La suma de los elementos de la diagonal secundaria no puede ser cero.

Tablas: Se muestra la tabla de contingencia de las variables pareadas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

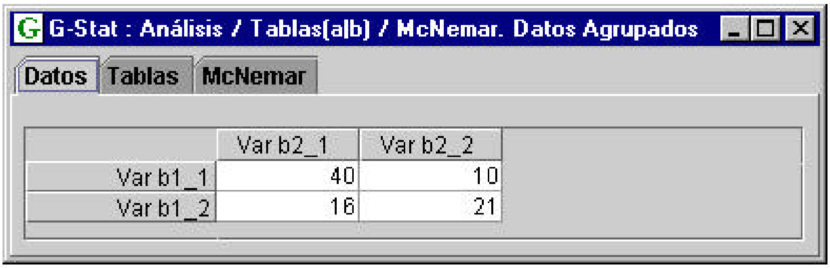
McNemar: Se presentan los resultados de la prueba de McNemar y Kappa.

Tablas (a|b) → McNemar. Datos Agrupados

A partir de datos agrupados se realiza la prueba McNemar. Adicionalmente, también se realiza el cálculo del coeficiente kappa de concordancia y de su significación estadística. Los datos agrupados se introducen directamente en una cuadrícula en forma de tabla de 2 filas y 2 columnas. Los fundamentos teóricos y la formulación son idénticos a los presentados en la opción de análisis anterior con datos a partir de un fichero.

Manejo del programa

Datos Agrupados: La disposición de la tabla para la entrada de datos es la siguiente:



En la pestaña de datos agrupados hay una cuadrícula que permite la entrada directa del número de casos dentro de cada posible combinación de categorías o niveles. Las variables deben ser dicotómicas, con lo que la tabla resultante es siempre dos por dos. Por defecto aparecen en la tabla unos valores que deben ser sustituidos por los datos del usuario.

No dejar en la tabla filas o columnas con todos los valores faltantes o con todos los valores iguales a cero. Las celdas de la tabla no admiten valores faltantes, negativos, decimales o alfanuméricos.

Se asume que la diagonal de concordancia viene dada por la casilla superior izquierda y la casilla inferior derecha. Si no es así, reordenar los datos para conseguir dicha disposición. La suma de los elementos de la diagonal secundaria no puede ser cero.

Tablas: Se muestra la tabla de contingencia de las variables pareadas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden obtener los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

McNemar: Se presentan los resultados de la prueba de McNemar y Kappa.

Epidemiología (b|b)

Contiene un submenú con técnicas de Epidemiología, con las opciones de Tablas, Tablas (Datos Agrupados) y Mantel-Haenszel (Datos Agrupados).

Epidemiología (b|b) → Tablas

Las principales medidas de efecto para variables dicotómicas en Epidemiología son el riesgo relativo RR y el odds ratio OR.

El riesgo relativo es válido generalmente en estudios de cohortes, mientras que el odds ratio lo es tanto en estudios de cohortes como en estudios de casos y controles. La interpretación para las dos medidas de efecto es similar:

- el valor uno indica ausencia de efecto
- valores superiores a uno, factor de riesgo
- valores inferiores a uno, factor protector

La notación que se utiliza es la siguiente:

	Factor de exposición		
Enfermedad	Sí	No	Total
Sí	a	b	r_1
No	c	d	r_2
Total	c_1	c_2	n

En el caso de celdas vacías, se suma 0.5 en todas las celdas para evitar problemas de cálculo en las medidas de efecto.

El odds ratio en estudios prospectivos se define por:

$$OR = \frac{\dfrac{Prob(Enfermedad \mid Expuestos)}{1 - Prob(Enfermedad \mid Expuestos)}}{\dfrac{Prob(Enfermedad \mid No\ Expuestos)}{1 - Prob(Enfermedad \mid No\ Expuestos)}}$$

En estudios retrospectivos se define como:

$$OR = \frac{\dfrac{Prob(Expuestos \mid Enfermos)}{1 - Prob(Expuestos \mid Enfermos)}}{\dfrac{Prob(Expuestos \mid No\ Enfermos)}{1 - Prob(Expuestos \mid No\ Enfermos)}}$$

Se demuestra que ambas expresiones son equivalentes, por lo que la estimación del OR para estudios prospectivos y retrospectivos es la misma y viene dada por

$$OR = \frac{a \cdot d}{b \cdot c}$$

Para determinar el IC(OR) es conveniente trabajar con el transformado logarítmico del OR, para lo cual es necesario calcular

$$EE(LnOR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

con lo que:

$$IC95\%(OR) = \left(\exp(LnOR - 1.96 \cdot EE(LnOR))\ ;\ \exp(LnOR + 1.96 \cdot EE(LnOR))\right)$$

da el intervalo de confianza del OR directamente.

La significación del OR viene dada por la prueba Chi-Cuadrado que en el caso de una tabla 2x2 tiene la siguiente expresión:

$$\chi^2 = \frac{(a\,d - b\,c)^2\,n}{r_1 r_2 c_1 c_2}$$

que sigue una distribución χ^2 con (2 - 1)x(2 - 1) = 1 grado de libertad.

El riesgo relativo sólo tiene sentido calcularlo en estudios prospectivos, y se define como

$$RR = \frac{\text{Prob}(\text{Enfermedad} \mid \text{Expuestos})}{\text{Prob}(\text{Enfermedad} \mid \text{No Expuestos})}$$

y que se estima por

$$RR = \frac{a \cdot c_2}{b \cdot c_1}$$

Para determinar el IC(RR) es conveniente trabajar con el transformado logarítmico del RR, para lo cual es necesario calcular

$$EE(LnRR) = \sqrt{\frac{1}{a} - \frac{1}{c_1} + \frac{1}{b} - \frac{1}{c_2}}$$

con lo que

$$IC95\%(RR) = \left(\exp\left(LnRR - 1.96 \cdot EE\left(LnRR\right)\right) ; \exp\left(LnRR + 1.96 \cdot EE\ \left(LnRR\right)\right)\right)$$

da el intervalo de confianza del RR directamente.

La significación del RR es la misma que la significación del OR.

Manejo del programa

Variables: Se identifica la *Variable Respuesta (resolución)* que aparecerá en filas y la *Variable Explicativa (factor de exposición)* que aparecerá en columnas, ambas dicotómicas.

Tablas: Se muestra la tabla de contingencia de dos variables cualitativas dicotómicas. En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden dar los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Si la variable explicativa (factor de exposición) se localiza en la variable columna, los porcentajes por columnas corresponden a los porcentajes (riesgos) de los grupos expuestos y no expuestos.

RR y OR: Se presentan las principales medidas de efecto para variables dicotómicas en Epidemiología: el Riesgo Relativo RR y el Odds

Ratio OR. Adicionalmente se facilitan los intervalos de confianza para el riesgo relativo poblacional y el odds ratio poblacional. El nivel de confianza se puede modificar en la ventana de opciones considerándolo = (1-alfa).

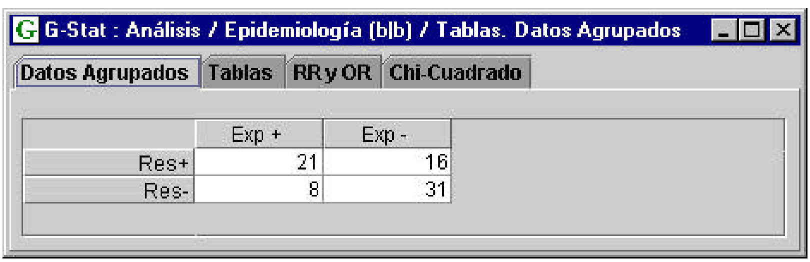
Chi-Cuadrado: Se contrasta la independencia entre la variable explicativa (factor de exposición) y la variable respuesta o resolución, mediante la prueba de Chi-Cuadrado. Se contrasta si los valores poblacionales de las medidas de efecto (RR y OR) son diferentes de 1.

Epidemiología (b|b) → Tablas. Datos Agrupados

A partir de datos agrupados se calculan las medidas de efecto básicas en Epidemiología. Los datos agrupados se introducen directamente en una cuadrícula en forma de tabla de 2 filas y 2 columnas. Los fundamentos teóricos y la formulación son idénticos a los presentados en la opción de análisis anterior con datos a partir de un fichero. En el caso de celdas vacías, se suma 0.5 en todas las celdas para evitar problemas de cálculo en las medidas de efecto.

Manejo del programa

Datos Agrupados: La disposición de la tabla para la entrada de datos es la siguiente:



La cuadrícula permite la entrada directa del número de casos dentro de cada posible combinación de categorías o niveles. Las variables deben ser dicotómicas, con lo que la tabla es dos por dos. La tabla está definida por las frecuencias de respuestas positivas y negativas para los grupos de expuestos y no expuestos. Por defecto aparecen en la tabla unos valores que deben ser sustituidos por los datos del usuario.

No dejar en la tabla filas o columnas con todos los valores faltantes o con todos los valores iguales a cero. Las celdas de la tabla no admiten valores faltantes, negativos, decimales o alfanuméricos.

Tablas: Se muestra la tabla de contingencia de las dos variables dicotómicas . En cada celda aparece en términos absolutos el número de individuos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden dar los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Si la variable explicativa (factor de exposición) se localiza en la variable columna, los porcentajes por columnas corresponden a los porcentajes (riesgos) de los grupos expuestos y no expuestos.

RR y OR: Se presentan las principales medidas de efecto para variables dicotómicas en Epidemiología: el Riesgo Relativo RR y el Odds Ratio OR. Adicionalmente se facilitan los intervalos de confianza para el riesgo relativo poblacional y el odds ratio poblacional. El nivel de confianza se puede modificar en la ventana de opciones considerándolo = (1-alfa).

Chi-Cuadrado: Se contrasta la independencia entre la variable explicativa (factor de exposición) y la variable respuesta o resolución, mediante la prueba de Chi-Cuadrado. Se contrasta si los valores poblacionales de las medidas de efecto (RR y OR) son diferentes de 1.

Epidemiología (b|b) → Mantel-Haenszel. Datos Agrupados (c|(b|b))

Realiza la prueba de Mantel-Haenszel que combina información en Epidemiología de varias tablas 2x2 que estudian el mismo factor de exposición y la misma resolución o respuesta. Esta técnica junto con la metodología Logit, son las técnicas básicas que se utilizan en Meta-Análisis.

La notación interna que utiliza el programa es la de k tablas 2x2 de la forma siguiente, con k el número de tablas o estudios (niveles) que hay que combinar, y j=1,…,k:

	Factor de exposición		
Respuesta	Sí	No	Total
Sí	a_j	b_j	r_{1j}
No	c_j	d_j	r_{2j}
Total	s_{1j}	s_{2j}	n_j

Esta notación interna se presenta en el interfaz de entrada de la forma:

	Nivel 1	Nivel 2	Nivel j	Nivel k
Res + / Exp+	a_1	a_2	a_j	a_k
Res+ / Exp-	b_1	b_2	b_j	b_k
Res- / Exp+	c_1	c_2	c_j	c_k
Res- / Exp-	d_1	d_2	d_j	d_k

Las fórmulas para el cálculo en cada nivel del Riesgo Relativo, Odds Ratio y Chi-2 de significación son las descritas en la opción anterior de Epidemiología. Las fórmulas del Riesgo Relativo Mantel-Haenszel, Odds Ratio Mantel-Haenszel y Chi-2 Mantel-Haenszel de significación se describen a continuación. Asimismo, se dan las fórmulas para calcular la homogeneidad de los estudios o niveles. Si se rechaza la hipótesis nula de homogeneidad (y se concluya heterogeneidad), los valores estimados por Mantel-Haenszel son cuestionables.

Riesgo Relativo Mantel-Haenszel

El riesgo relativo Mantel-Haenszel se calcula como

$$RR_{MH} = \frac{\sum_{j=1}^{k} \frac{a_j s_{2j}}{n_j}}{\sum_{j=1}^{k} \frac{b_j s_{1j}}{n_j}}$$

Para determinar el $IC(RR_{MH})$ es conveniente trabajar con el transformado logarítmico del RR_{MH}, para lo cual es necesario calcular

$$EE(LnRR_{MH}) = \sqrt{\frac{\sum_{j=1}^{k} \frac{r_{1j} s_{1j} s_{2j} - a_j b_j n_j}{n_j^2}}{\left[\sum_{j=1}^{k} \frac{a_j s_{2j}}{n_j}\right]\left[\sum_{j=1}^{k} \frac{b_j s_{1j}}{n_j}\right]}}$$

con lo que el IC95%(RR_{MH}) viene dado directamente por

$$\left(\exp\left(LnRR_{MH} - 1.96 \cdot EE\left(LnRR_{MH}\right)\right) ; \exp\left(LnRR_{MH} + 1.96 \cdot EE\left(LnRR_{MH}\right)\right)\right)$$

Odds Ratio Mantel-Haenszel

El odds ratio Mantel-Haenszel se calcula como

$$OR_{MH} = \frac{\sum_{j=1}^{k} \frac{a_j d_j}{n_j}}{\sum_{j=1}^{k} \frac{b_j c_j}{n_j}}$$

Para determinar el IC(OR_{MH}) es conveniente trabajar con el transformado logarítmico del RR_{MH}, para lo cual es necesario calcular

$$EE(LnOR_{MH}) = \sqrt{Aux_1 + Aux_2 + Aux_3}$$

siendo

$$Aux_1 = \frac{\sum_{j=1}^{k} \left(\frac{a_j + d_j}{n_j}\right)\left(\frac{a_j d_j}{n_j}\right)}{2\left[\sum_{j=1}^{k} \frac{a_j d_j}{n_j}\right]^2}$$

$$Aux_2 = \frac{\sum_{j=1}^{k} \left(\frac{a_j + d_j}{n_j}\right)\left(\frac{b_j c_j}{n_j}\right) + \left(\frac{b_j + c_j}{n_j}\right)\left(\frac{a_j d_j}{n_j}\right)}{2\left[\sum_{j=1}^{k} \frac{a_j d_j}{n_j}\right]\left[\sum_{j=1}^{k} \frac{b_j c_j}{n_j}\right]}$$

$$Aux_3 = \frac{\sum_{j=1}^{k} \left(\frac{b_j + c_j}{n_j}\right)\left(\frac{b_j c_j}{n_j}\right)}{2\left[\sum_{j=1}^{k} \frac{b_j c_j}{n_j}\right]^2}$$

con lo que el IC95%(OR_{MH}) viene dado directamente por

$$\left(\exp\left(LnOR_{MH} - 1.96 \cdot EE\left(LnOR_{MH}\right)\right); \exp\left(LnOR_{MH} + 1.96 \cdot EE\left(LnOR_{MH}\right)\right)\right)$$

Homogeneidad de Riesgos Relativos

Se construyen k tablas 2x2 ficticias que tengan por RR_j el mismo RR_{MH} en cada estudio o nivel j, j=1,…,k, de la forma siguiente:

	Factor de exposición		
Respuesta	Sí	No	Total
Sí	a^*_j	b^*_j	r_{1j}
No	c^*_j	d^*_j	r_{2j}
Total	s_{1j}	s_{2j}	n_j

con

$$RR_{MH} = \frac{a_j^* s_{2j}}{b_j^* s_{1j}}$$

Se plantea una ecuación lineal en a^*_j y de dicha ecuación se resuelven el valor del resto de las celdas:

$$RR_{MH} = \frac{a_j^* s_{2j}}{\left(r_{1j} - a_j^*\right) s_{1j}}$$

$$a_j^* = \frac{r_{1j} s_{1j} RR_{MH}}{s_{2j} + RR_{MH} s_{1j}}$$

$$b_j^* = r_{1j} - a_j^*$$

$$c_j^* = s_{1j} - a_j^*$$

$$d_j^* = r_{2j} - \left(s_{1j} - a_j^*\right)$$

A partir de estos valores se construye el siguiente estadístico de homogeneidad

$$\sum_{j=1}^{k} \left[\frac{\left(a_j - a_j^*\right)^2}{a_j^*} + \frac{\left(b_j - b_j^*\right)^2}{b_j^*} + \frac{\left(c_j - c_j^*\right)^2}{c_j^*} + \frac{\left(d_j - d_j^*\right)^2}{d_j^*} \right]$$

Este estadístico se ha definido de forma análoga al estadístico de homogeneidad de Breslow-Day para odds ratio y está en fase experimental de estudio, pero siguiendo la analogía con Breslow-Day, seguirá una distribución chi-cuadrado con 1 grado de libertad.

Este estadístico de homogeneidad puede dar valores negativos, en esos caso el programa escribe en los resultados “No Aplicable”.

Homogeneidad Breslow-Day de Odds Ratios

Se construyen k tablas 2x2 ficticias que tengan por OR_j el mismo OR_{MH} en cada estudio o nivel j, j=1,...,k, de la forma siguiente

	Factor de exposición		
Respuesta	Sí	No	Total
Sí	a^*_j	b^*_j	r_{1j}
No	c^*_j	d^*_j	r_{2j}
Total	s_{1j}	s_{2j}	n_j

con

$$OR_{MH} = \frac{a_j^* d_j^*}{b_j^* c_j^*}$$

Se plantea una ecuación cuadrática en a^*_j y de dicha ecuación se resuelven el valor del resto de las celdas:

$$OR_{MH} = \frac{a_j^*\left[r_{2j} - \left(s_{1j} - a_j^*\right)\right]}{\left[r_{1j} - a_j^*\right]\left[s_{1j} - a_j^*\right]}$$

$$a_j^{*2}\left(OR_{MH} - 1\right) + a_j^*\left(s_{1j} - r_{2j} - s_{1j}OR_{MH} - r_{1j}OR_{MH}\right) + OR_{MH}r_{1j}s_{1j} = 0$$

$$a_j^* = \frac{-\left(s_{1j} - r_{2j} - s_{1j}OR_{MH} - r_{1j}OR_{MH}\right) \pm \sqrt{AUX_1}}{2\left(OR_{MH} - 1\right)}$$

$$AUX_1 = \left(s_{1j} - r_{2j} - s_{1j}OR_{MH} - r_{1j}OR_{MH}\right)^2 - 4\left(OR_{MH} - 1\right)OR_{MH}r_{1j}s_{1j}$$

$$b_j^* = r_{1j} - a_j^*$$

$$c_j^* = s_{1j} - a_j^*$$

$$d_j^* = r_{2j} - \left(s_{1j} - a_j^*\right)$$

La solución de la anterior ecuación para a^*_j será aquella que verifique $a^*_j \geq 0$, $a^*_j \leq r_{1j}$ y $a^*_j \leq s_{1j}$.

A partir de estos valores se construye el estadístico Breslow-Day de homogeneidad

$$\sum_{j=1}^{k}\left[\frac{\left(a_j - a_j^*\right)^2}{a_j^*} + \frac{\left(b_j - b_j^*\right)^2}{b_j^*} + \frac{\left(c_j - c_j^*\right)^2}{c_j^*} + \frac{\left(d_j - d_j^*\right)^2}{d_j^*}\right]$$

Este estadístico sigue una distribución chi-cuadrado con 1 grado de libertad.

Chi-Cuadrado Mantel-Haenszel

El estadístico Chi-Cuadrado Mantel-Haenszel se calcula como

$$\chi^2_{MH} = \frac{\left[\sum_{j=1}^{k}\left(a_j - E(a_j)\right)\right]^2}{\sum_{j=1}^{k} V[a_j]}$$

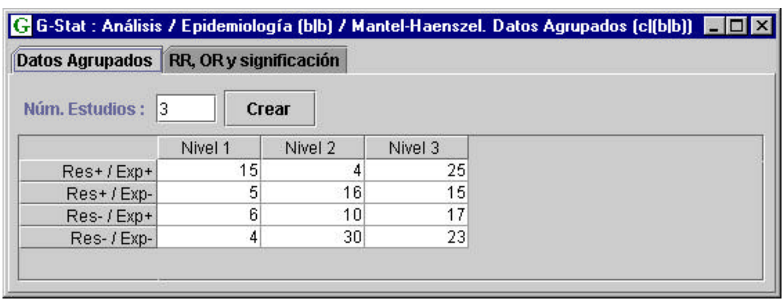
que sigue una distribución Chi-Cuadrado con un grado de libertad, siendo

$$E[a_j] = \frac{r_{1j}s_{1j}}{n_j}$$

$$V[a_j] = \frac{r_{1j}r_{2j}s_{1j}s_{2j}}{n_j^2\left(n_j - 1\right)}$$

Manejo del programa

Datos Agrupados: La disposición de la tabla para la entrada de datos es la siguiente:



La cuadrícula permite la entrada directa de las frecuencias de respuestas positivas y negativas para los grupos de expuestos y no expuestos. Se puede definir el número de estudios o niveles de la variable de estratificación. Cada columna está asociada a un estudio. El botón "Crear" prepara la estructura de la tabla ajustada al número de estudios. Por defecto aparecen en la tabla unos valores que deben ser sustituidos por los datos del usuario.

No dejar en la tabla filas o columnas con todos los valores faltantes o con todos los valores iguales a cero. Las celdas de la tabla no admiten valores faltantes, negativos, decimales o alfanuméricos.

RR, OR y significación: Se presenta para cada nivel de la variable formadora de capas, las medidas de efecto RR y OR, así como su intervalo de confianza y su significación global.

Diagnóstico (b|b)

Contiene un submenú con técnicas de Diagnóstico, con las opciones de Tablas, Tablas (Datos Agrupados) y curvas ROC (“Receiver Operating Characteristic curves”).

Diagnóstico (b|b) → Tablas

En esta opción se describen las técnicas de diagnóstico que intentan estudiar el comportamiento de un test o prueba diagnóstica (que se considera como variable explicativa) en relación a si pronostica bien o mal, en un sujeto, la presencia o ausencia de una “enfermedad” (que se considera como variable respuesta). Esta variable que se intenta pronosticar se conoce también como “Estado de la naturaleza (EN)”, que representa la verdadera condición del sujeto. Se supone que el test o prueba diagnóstico, al igual que el EN, tiene como posibles resultados el valor positivo y negativo.

La notación que se utiliza es la siguiente:

EN	Test +	Test -	Total
+	a	b	r_1
-	c	d	r_2
Total	c_1	c_2	n

Los valores a y d representan a aquellos individuos que son pronosticados correctamente. Los índices de fiabilidad diagnóstica que se calculan son: sensibilidad, especificidad, valores predictivos de ocurrencia para cada resultado del test, likelihood ratio y odds ratio.

En estudios transversales (donde no se prefija ningún marginal r_1, r_2, c_1 ó c_2, sólo se prefija n) se puede calcular la prevalencia muestral, que viene dada por

$$\text{Prevalencia} = \frac{r_1}{n}$$

Algunos de los índices de fiabilidad diagnóstica no dependen de la prevalencia teórica, otros, sin embargo sí. En los casos en que sí dependa y se esté en un estudio transversal, es necesario tener una estimación (por otros estudios, referencias bibliográfica) de la prevalencia teórica. Esta estimación se denota por Prev.

Sensibilidad y especificidad

La sensibilidad y la especificidad de un test se calculan como

$$\text{Sens} = \frac{a}{r_1}$$

$$\text{Espe} = \frac{d}{r_2}$$

La sensibilidad representa la probabilidad de test positivo en el grupo EN positivo (acierto en “enfermos”) y la sensibilidad la probabilidad de test negativo en el grupo EN negativo (acierto en “sanos”). La sensibilidad y la especificidad no dependen de la prevalencia.

El cálculo de los intervalos de confianza de la sensibilidad y de la especificidad se realizan teniendo en cuenta que son proporciones, por lo que se utilizan las fórmulas descritas en la opción Análisis / Cualitativa / Una Proporción.

Valores predictivos

Los valores predictivos de “enfermedad” de un test se calculan como

$$VP+ = \frac{a}{c_1}$$

$$VP- = \frac{b}{c_2}$$

El VP+ representa la Prob(EN positivo | test positivo), es decir la probabilidad de EN positivo en el grupo de test positivo. El VP- representa la Prob(EN positivo | test negativo), es decir la probabilidad de EN positivo en el grupo de test negativo. Los valores predictivos de “enfermedad” dependen de la prevalencia. Esto quiere decir que en estudios transversales las fórmulas anteriores son correctas para el cálculo de VP+ y VP-. En el caso de estudios no transversales es necesario calcularlos a partir de Prev y de los Likelihood ratio, según se verá a continuación.

El cálculo de los intervalos de confianza de los valores predictivos se dará en el caso general a partir de Prev y de los Likelihood ratio.

Likelihood ratio

Los Likelihood ratio de “enfermedad” de un test se calculan como

$$L(+) = \frac{Sens}{1 - Espe}$$

$$L(-) = \frac{1 - Sens}{Espe}$$

Los Likelihood ratio de “enfermedad” no dependen de la prevalencia.

Para determinar el IC(L(+)) es conveniente trabajar con el transformado logarítmico del L(+), para lo cual es necesario calcular

$$EE(LnL(+)) = \sqrt{\frac{1 - Sens}{a} + \frac{Espe}{c}}$$

con lo que el IC95%(L(+)) viene dado directamente por

$$\left(\exp\left(LnL(+) - 1.96 \cdot EE\left(LnL(+)\right)\right) ; \exp\left(LnL(+) + 1.96 \cdot EE\left(LnL(+)\right)\right)\right)$$

Para determinar el IC(L(-)) es conveniente trabajar con el transformado logarítmico del L(-), para lo cual es necesario calcular

$$EE(LnL(-)) = \sqrt{\frac{Sens}{b} + \frac{1-Espe}{d}}$$

con lo que el IC95%(L(-)) viene dado directamente por

$$\left(\exp\left(LnL(-) - 1.96 \cdot EE\left(LnL(-)\right)\right) ; \exp\left(LnL(-) + 1.96 \cdot EE\left(LnL(-)\right)\right)\right)$$

A partir de los Likelihood ratio y de Prev se pueden calcular VP+ y VP- en estudios no transversales, de la forma siguiente:

$$VP+ = Prob(EN+ \mid Test+) = \frac{Prev \cdot L(+)}{Prev \cdot L(+) + 1 - Prev}$$

$$VP- = Prob(EN+ \mid Test-) = \frac{Prev \cdot L(-)}{Prev \cdot L(-) + 1 - Prev}$$

Para determinar el IC(VP+) es necesario calcular

$$EE(VP+) = (VP+)(1 - VP+)\sqrt{\left(\frac{EE(Sens)}{Sens}\right)^2 + \left(\frac{EE(Ines)}{Ines}\right)^2 + \left(\frac{EE(Prev)}{Prev(1-Prev)}\right)^2}$$

siendo

$$Ines = 1 - Espe$$

$$EE(Sens) = \sqrt{\frac{\frac{a}{r_1}\left(1 - \frac{a}{r_1}\right)}{r_1}}$$

$$EE(Ines) = \sqrt{\frac{\left(1 - \frac{d}{r_2}\right)\frac{d}{r_2}}{r_2}}$$

$$EE(Prev) = \sqrt{\frac{Prev(1-Prev)}{n}}$$

con lo que el IC95%(VP+) viene dado por

$$\left((VP+) - 1.96 \cdot EE(VP+) ; (VP+) + 1.96 \cdot EE(VP+)\right)$$

Para determinar el IC(VP-) es necesario calcular

$$EE(VP-) = (VP-)(1 - VP-)\sqrt{\left(\frac{EE(Sens)}{Sens}\right)^2 + \left(\frac{EE(Ines)}{Ines}\right)^2 + \left(\frac{EE(Prev)}{Prev(1-Prev)}\right)^2}$$

con lo que el IC95%(VP-) viene dado por

$$\left((VP-) - 1.96 \cdot EE\left(VP-\right) \,;\, (VP-) + 1.96 \cdot EE\left(VP-\right)\right)$$

Odds ratio

El cálculo del odds ratio y de su intervalo de confianza se realiza según las fórmulas dadas en la opción de Epidemiología. Observar, no obstante, que

$$OR = \frac{L(+)}{L(-)}$$

Manejo del programa

Variables: Se identifica la *Variable respuesta* "Estado de la naturaleza" (EN*)* o verdadera condición del sujeto en filas y la *Variable explicativa* *"*resultado del Test" (prueba diagnóstica) en columnas, ambas dicotómicas.

En las pestañas posteriores de Tablas y Diagnóstico habrá que comprobar que la diagonal de concordancia viene dada por la casilla superior izquierda y la casilla inferior derecha y que, adicionalmente, los resultados positivos del Estado de la naturaleza y del Test deben estar situados en la casilla superior izquierda. Si no es así, ordenar los datos para conseguir dicha disposición.

Tablas: Se muestra la tabla de contingencia dos por dos de dos variables cualitativas dicotómicas. En cada celda aparece en términos absolutos el número de casos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden dar los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Diagnóstico: Aparecen los índices de fiabilidad diagnóstica, sensibilidad, especificidad, valores predictivos de ocurrencia para cada resultado del test, likelihood ratio y odds ratio. Asimismo, se presenta la prevalencia en la muestra.

La prevalencia se puede modelizar en las opciones. Los valores predictivos dependen del valor de la prevalencia muestral, la especificidad y la sensibilidad no.

Se incluyen los intervalos de confianza. En las opciones, el valor por defecto de alfa es 5% que corresponde a un IC del 95%.

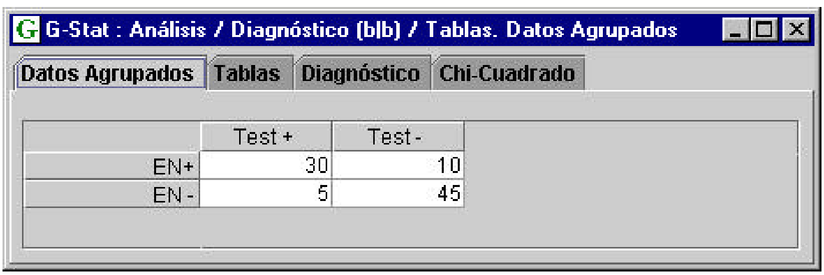
Chi-Cuadrado: Se contrasta la independencia entre el resultado del Test y la variable Estado de la naturaleza EN, mediante la prueba de Chi-Cuadrado. Este tipo de contraste no suele utilizarse en diagnóstico puesto que normalmente, aun siendo estadísticamente significativo el resultado, muchas veces los índices o medidas de fiabilidad diagnóstica no son suficientemente adecuados.

Diagnóstico (b|b) → Tablas. Datos Agrupados

A partir de datos agrupados se calculan las técnicas estadísticas básicas utilizadas en Diagnóstico. Los datos agrupados se introducen directamente en una cuadrícula en forma de tabla de 2 filas y 2 columnas. Los fundamentos teóricos y la formulación son idénticos a los presentados en la opción de análisis anterior con datos a partir de un fichero.

Manejo del programa

Datos Agrupados: La disposición de la tabla para la entrada de datos es la siguiente:



La cuadrícula permite la entrada directa del número de casos dentro de cada posible combinación de categorías o niveles. Las variables deben ser dicotómicas, con lo que la tabla es de dos por dos. La tabla está definida por las frecuencias de resultados del Test positivos y negativos para los dos Estados de la naturaleza. Por defecto aparecen en la tabla unos valores que deben ser sustituidos por los datos del usuario.

Se asume que la diagonal de concordancia viene dada por la casilla superior izquierda y la casilla inferior derecha. Adicionalmente, los resultados positivos del Estado de la naturaleza y del Test deben estar situados en la casilla superior izquierda. Si no es así, reordenar los datos para conseguir dicha disposición.

No dejar en la tabla filas o columnas con todos los valores faltantes o con todos los valores iguales a cero. Las celdas de la tabla no admiten valores faltantes, negativos, decimales o alfanuméricos. La suma de frecuencias de la diagonal de discordancias no debería ser cero.

Tablas: Se muestra la tabla de contingencia dos por dos de dos variables cualitativas dicotómicas. En cada celda aparece en términos absolutos el número de casos dentro de cada posible combinación de categorías.

Adicionalmente, se pueden dar los porcentajes de dichos valores con relación al número total de individuos en una categoría en fila (porcentajes fila), en una categoría en columna (porcentajes columna) y del total de individuos (porcentaje total). El programa calcula, por defecto, los porcentajes referidos al total de la tabla.

Diagnóstico: Aparecen los índices de fiabilidad diagnóstica, sensibilidad, especificidad, valores predictivos de ocurrencia para cada resultado del test, likelihood ratio y odds ratio. Asimismo, se presenta la prevalencia en la muestra.

La prevalencia se puede modelizar en las opciones. Los valores predictivos dependen del valor de la prevalencia muestral, la especificidad y la sensibilidad no dependen.

Se incluyen los intervalos de confianza. En las opciones, el valor por defecto de alfa es 5% que corresponde a un IC del 95%.

Chi-Cuadrado: Se contrasta la independencia entre el resultado del Test y la variable Estado de la naturaleza EN, mediante la prueba de Chi-Cuadrado. Este tipo de contraste no suele utilizarse en diagnóstico puesto que normalmente, aun siendo estadísticamente significativo el resultado, muchas veces los índices o medidas de fiabilidad diagnóstica no son suficientemente adecuadas.

Diagnóstico (b|b) → ROC (y|b)

Realiza el cálculo de las curvas ROC o curvas de rendimiento diagnóstico, que son técnicas de diagnóstico que intentan estudiar el comportamiento de un test o prueba diagnóstica cuantitativa (que se considera como variable explicativa) en relación a si pronostica bien o mal, en un sujeto, la presencia o ausencia de una "enfermedad" (que se considera como variable respuesta). Esta variable que se intenta pronosticar se conoce también como "Estado de la naturaleza (EN)", que representa la verdadera condición del sujeto. A diferencia de las opciones anteriores de Diagnóstico, se supone que el test o prueba diagnóstico es cuantitativa en vez de dicotómica. El EN sigue siendo dicotómica.

Esta técnica se basa en discretizar la variable explicativa en tramos. Para ello se ordenan de menor a mayor los distintos valores que toma la variable explicativa, que se denotan por x_1, x_2, ..., x_k. Los tramos que se consideran son los dados por los puntos de corte:

$$<x_1,\ (x_1+x_2)/2,\ ...,\ (x_{k-1}+x_k)/2,\ >x_k$$

En cada punto de corte p_c se construye una tabla 2x2 de la forma siguiente:

	Test		
EN	Test>=p_c	Test<p_c	Total
+	a_j	b_j	r_{1j}
-	c_j	d_j	r_{2j}
Total	c_{1j}	c_{2j}	n_j

donde Test>=p_c se interpreta como un test que da resultado positivo y Test<p_c se interpreta como un test que da resultado negativo. Se tiene, por tanto, que

a_j: Verdaderos positivos, EN+ y Test+

b_j: Falsos negativos, EN+ y Test -

c_j: Falsos positivos, EN- y Test+

d_j: Verdaderos negativos, EN- y Test -

En cada una de estas tablas se calcula la Sens_j y la $(1\text{-Espe})_j$ con j=1, ...,k. La curva ROC es la representación gráfica de estos valores obtenidos, donde en el eje horizontal se representa 1-Especificidad y en el eje vertical la Sensibilidad.

El Área bajo la curva (ABC) de esta curva ROC da una medida del grado de rentabilidad diagnóstica del Test. El ABC varía entre 0 y 1. Valores cercanos a 0.5 implican ausencia de rentabilidad diagnóstica. Valores menores de 0.5 del ABC indican diagnósticos inversos.

El área bajo la curva ROC se calcula por el método trapezoidal de la forma siguiente:

$$\sum_{j=1}^{k}\left((1-\text{Espe})_{j}-(1-\text{Espe})_{j-1}\right)\left(\text{Sens}_{j-1}+\frac{\text{Sens}_{j}-\text{Sens}_{j-1}}{2}\right)$$

El punto de probabilidad de corte óptimo para a partir del valor de la variable explicativa determinar la ocurrencia de la variable respuesta, será aquel en donde la curva más se acerca a la esquina superior izquierda del gráfico, que es el punto con Sensibilidad y Especificidad igual a 1.

Manejo del programa

Variables: Se identifica la *Variable respuesta* dicotómica "Estado de la naturaleza" (EN) o verdadera condición del sujeto y la *Variable explicativa* cuantitativa “resultado del Test” (prueba diagnóstica).

ROC Tabla: Para la variable cuantitativa Test se generan sucesivos puntos de corte. Una vez dicotomizada la variable explicativa cuantitativa se calculan los índices diagnósticos Sensibilidad y 1-Especificidad para los diferentes cortes. El ABC se estima por el método trapezoidal, indicándose las Areas bajo la Curva parciales. El ABC global se presenta al final de las ABC parciales y en la pestaña ROC Gráfico.

En opciones se identifica el código del EN relacionado con el valor positivo, generalmente "enfermedad”.

ROC Gráfico: Se muestra la curva ROC que es la representación gráfica de los valores obtenidos según se indica en la pestaña ROC Tabla, donde en el eje horizontal se representa 1-Especificidad y en el eje vertical la Sensibilidad.

En opciones se identifica el código del EN relacionado con el valor positivo, generalmente "enfermedad", los títulos y los valores mínimo y máximo de los ejes. Estas opciones son independientes de las de la pestaña ROC Tabla y deberán igualarse para obtener resultados compatibles.

Dos Grupos (b|y)

Contiene un submenú con diferentes técnicas estadísticas para una variable explicativa dicotómica o binaria (b) que intenta explicar una variable explicativa cuantitativa (y). Se incluyen las siguientes pruebas**:** t-Student, t-Student para datos pareados, F-Snedecor, t-Student y F-Snedecor para datos agrupados, Mann-Whitney (Wilcoxon), Wilcoxon para datos pareados y Signos para datos pareados.

Dos Grupos (b|y) → t-Student

Realiza el intervalo de confianza del parámetro diferencia poblacional de dos medias y un contraste de hipótesis de dos medias mediante la prueba t-Student.

Para calcular el IC(1 - α)% para la diferencia de medias suponiendo igualdad de varianzas, se necesita calcular el error estándar de la diferencia de medias que, en el supuesto de igualdad de varianzas, tiene la expresión

$$EE(\bar{x}_1 - \bar{x}_2) = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

siendo $\bar{x}_1$ y $\bar{x}_2$ las medias por cada grupo y s^2 la varianza conjunta (“pooled variance”), que tiene por expresión

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

siendo s_1^2 y s_2^2 las varianzas muestrales para cada grupo. En segundo lugar para calcular el IC deseado se necesita el valor de la t-Student $t_{1-\alpha/2;gl}$ con grados de libertad gl = (n_1 - 1) + (n_2 – 1) = (n_1 + n_2 – 2) , con lo que

$$IC(1 - \alpha)\%(\bar{x}_1 - \bar{x}_2) = \left[(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2,gl} EE(\bar{x}_1 - \bar{x}_2)\right]$$

proporciona el IC buscado.

Para calcular el IC(1 - α)% para la diferencia de medias suponiendo no igualdad de varianzas, se necesita calcular el error estándar de la diferencia de medias que, en el supuesto de no igualdad de varianzas, tiene la expresión

$$EE(\overline{x}_1 - \overline{x}_2) = \sqrt{EE(\overline{x}_1)^2 + EE(\overline{x}_2)^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

En segundo lugar, para calcular el IC deseado se necesita el valor de la t-Student $t_{1-\alpha/2;gl}$ con grados de libertad gl dados por la siguiente expresión, llamada de Satterthwaite

$$gl = \frac{\left[EE(\overline{x}_1 - \overline{x}_2)\right]^4}{\frac{1}{n_1 - 1}\left[EE(\overline{x}_1)\right]^4 + \frac{1}{n_2 - 1}\left[EE(\overline{x}_2)\right]^4}$$

con lo que

$$IC(1-\alpha)\%(\overline{x}_1 - \overline{x}_2) = \left[(\overline{x}_1 - \overline{x}_2) \pm t_{1-\alpha/2,gl} EE(\overline{x}_1 - \overline{x}_2)\right]$$

proporciona el IC buscado.

Para llevar a cabo el contraste

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

suponiendo igualdad de varianzas poblacionales, se construye el estadístico de contraste experimental t dado por

$$t = \frac{\overline{x}_1 - \overline{x}_2}{EE(\overline{x}_1 - \overline{x}_2)} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

que bajo la hipótesis nula sigue una distribución t-Student con grados de libertad gl = $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$.

Para llevar a cabo el contraste

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

suponiendo no igualdad de varianzas poblacionales, se construye el estadístico de contraste experimental t dado por

$$t = \frac{\overline{x}_1 - \overline{x}_2}{EE(\overline{x}_1 - \overline{x}_2)} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

que bajo la hipótesis nula sigue una distribución t-Student con grados de libertad gl de Satterthwaite.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* dicotómica que forma los dos grupos. En los dos grupos, la variable respuesta debe tener desviación típica mayor que cero.

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Cajas: Gráfico de dos Cajas de la variable respuesta para los dos valores de la variable explicativa. Ver menú Gráficos.

t-Student: Se presenta el intervalo de confianza del parámetro diferencia poblacional de dos medias y los resultados del contraste de hipótesis de dos medias mediante la prueba t-Student.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

- El programa asume, por defecto, igualdad de varianzas entre las variables, pero puede ser modificado.

Establecer si hay diferencia de Edad en relación a los Tratamientos.

Resultados de la prueba t-Student asumiendo igualdad de varianzas.

```
Estimación y Contraste de Dos Medias Poblacionales de EDAD por FARMACO
==========================================================================

Variable Respuesta:       EDAD
Variable Explicativa:     FARMACO

 Grupo                    1          2
----------------------------------------------
 Tamaños Muestrales       20         20
 Medias:                  22.6000    22.2300
 Desviaciones Típicas:    1.6588     1.1921
 E. E. de las Medias:     0.3709     0.2666
----------------------------------------------

Varianza Conjunta:                2.0864
E. E. de la Diferencia de Medias: 0.4568

Grados de Libertad:           38.0000
Diferencia de Medias          0.3700

Estimación
----------
I.C.  al 95.00% para la diferencia de medias: 0.3700 +/- 0.9247   [-0.5547, 1.2947]

t-Student
---------
Hipótesis Nula:                diferencia de medias = 0.0000
Hipótesis Alternativa:         no igual
t-Student:                     0.8100
p-valor:                       0.4230
```

Dos Grupos (b|y) → t-Student. Pareados

Realiza el intervalo de confianza del parámetro media poblacional de la diferencia de dos variables pareadas y un contraste de hipótesis de una media mediante la prueba t-Student para dos variables pareadas.

La prueba t-Student para datos pareados calcula en primer lugar una nueva variable Vardif que es igual a la primera variable Var1 menos la segunda variable Var2. Para realizar el contraste

$$H_0\text{: media(Vardif)} = \mu_0$$

$$H_1\text{: media(Vardif)} \neq \mu_0$$

se calcula el estadístico de contraste t como

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s}{n}}}$$

donde es la media y s la desviación típica de la nueva variable, que sigue una distribución t-Student con gl = n - 1 grados de libertad.

Manejo del programa

Variables: Se identifica la *1ª Variable pareada* y la *2ª Variable pareada*, ambas cuantitativas y expresadas en las mismas unidades. El programa operará internamente con la diferencia de las dos variables. La variable diferencia debe tener desviación típica mayor que cero.

Estadísticos: Estadísticos de la diferencia de las dos variables pareadas. Se detallan en Cuantitativa (y).

Cajas: Gráfico de Cajas para la diferencia de las dos variables pareadas. Ver menú Gráficos.

t-Student. Pareados: Se presenta el intervalo de confianza del parámetro media poblacional de la diferencia de dos variables pareadas y los resultados del contraste de hipótesis de una media mediante la prueba t-Student para dos variables pareadas.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Contrastar si la media de FC2 menos FC1 es superior a 60.

Resultados descriptivos de la prueba t-Student para datos pareados.

```
Estadísticos para la variable FC2-FC1
==========================================================================


-------------------------------------
 Estadístico               FC2-FC1
-------------------------------------
 N                         40
 Media                     61.7500
 Mediana                   61.0000
 Desviación Típica         10.9772
 Mínimo                    42.0000
 Máximo                    82.0000
 Asimetría Estandarizada   -0.0433
 Curtosis Estandarizada    -1.0885
-------------------------------------
```

Resultados de la prueba t-Student para datos pareados.

```
Estimación y Contraste de Una Media Poblacional para FC2 menos FC1
========================================================================

Tamaño Muestral:   40
Media:             61.7500

Estimación
----------
I.C. inferior al 95.00% para la media:  61.7500 - 2.9244  [58.8256]

t-Student
---------
Hipótesis Nula:                 media = 60.0000
Hipótesis Alternativa:          mayor que
Estadístico de contraste t:     1.0083
p-valor:                        0.1598
```

Dos Grupos (b|y) → F-Snedecor

Realiza el intervalo de confianza del parámetro cociente poblacional de dos varianzas y un contraste de hipótesis de dos varianzas mediante la prueba F-Snedecor.

La expresión para calcular el IC(1 - α)% para el cociente de varianzas es

$$IC95\%\left(\frac{\sigma_1^2}{\sigma_2^2}\right)=\left(\frac{s_1^2/s_2^2}{F_{1-\alpha/2;gln;gld}}\ ;\ \frac{s_1^2}{s_2^2}F_{1-\alpha/2;gld;gln}\right)$$

siendo s_1^2 y s_2^2 las varianzas muestrales para cada grupo donde $F_{1-\alpha/2;\ gln;\ gld}$ se calcula a partir de una F-Snedecor siendo gln los grados de libertad del numerador, que se calculan como el tamaño muestral del grupo con mayor varianza muestral menos uno, y gld los grados de libertad del denominador que se calculan como el tamaño muestral del grupo con menor varianza muestral menos uno.

Para llevar a cabo el contraste

$$H_0: \sigma_1 - \sigma_2 = 0$$

$$H_1: \sigma_1 - \sigma_2 \neq 0$$

mediante la prueba F-Snedecor de comparación de varianzas se construye el estadístico de contraste experimental F dado por

$$F=\frac{\text{máx}\left\{s_1^2\,;\,s_2^2\right\}}{\text{mín}\left\{s_1^2\,;\,s_2^2\right\}}$$

que bajo la hipótesis nula sigue una distribución F-Snedecor siendo gln los grados de libertad del numerador y gld los grados de libertad del denominador. En el caso de no poder rechazar la hipótesis nula (p-valor > 0.05) se considera que las dos varianzas son iguales (homogéneas).

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* dicotómica que forma los dos grupos. En los dos grupos, la variable respuesta debe tener desviación típica mayor que cero.

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Cajas: Gráfico de dos Cajas de la variable respuesta para los dos valores de la variable explicativa. Ver menú Gráficos.

F-Snedecor: Se presenta el intervalo de confianza del parámetro cociente poblacional de dos varianzas y los resultados del contraste de hipótesis de dos varianzas mediante la prueba F-Snedecor.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 1, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Se desea comprobar si el cociente de varianzas es igual a 1 para la variable respuesta FC2FC1, con Farmaco como variable explicativa.

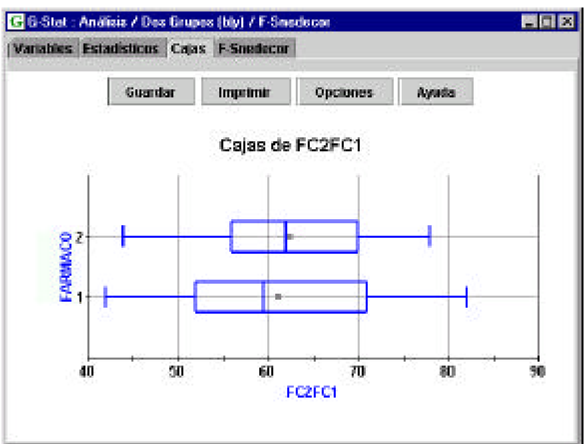


Gráfico de cajas de la opción F-Snedecor.

Resultados de la prueba F-Snedecor.

```
Estimación y Contraste de Dos Varianzas Poblacionales de FC2FC1 por FARMACO
==========================================================================
Variable Respuesta:          FC2FC1
Variable Explicativa:        FARMACO

 Grupo                     1           2
-----------------------------------------------
 Tamaños Muestrales       20          20
 Desviaciones Típicas:    12.4108     9.6206
 Varianzas:               154.0289    92.5553
-----------------------------------------------

Cociente de Varianzas:       1.6642

Estimación
----------
I.C.  al 95.00% para el cociente de varianzas:  [0.6587, 4.2045]

F-Snedecor
----------
Hipótesis Nula:                cociente de varianzas= 1.0000
Hipótesis Alternativa:         no igual
Estadístico de contraste F:    1.6642
p-valor:                       0.2758
```

Dos Grupos (b|y) → t-Student y F-Snedecor. Datos Agrupados

A partir de los datos resumidos de tamaño muestral, media y desviación típica de cada uno de los grupos, se puede realizar, sin el fichero de datos, las opciones Análisis / 2 Grupos (b|y) / t-Student y Análisis / 2 Grupos (b|y) / F-Snedecor. Los fundamentos teóricos y la formulación son idénticos a los presentados en las opciones anteriores respectivas con datos a partir de fichero.

Manejo del programa

Datos Agrupados:

Los datos necesarios son:

- Nombre Grupo 1: Nombre Categoría1.
- Nombre Grupo 2: Nombre Categoría2.

- Tamaño Grupo 1: Valor n1.
- Tamaño Grupo 2: Valor n2.
- Media Grupo 1: Valor de media1.
- Media Grupo 2: Valor de media2.
- Desv. Típica Grupo 1: Valor de s1.
- Desv. Típica Grupo 2: Valor de s2.

Los dos tamaños muestrales tienen que ser mayores que cero. En los dos grupos la variable respuesta debe tener desviación típica mayor que cero.

t-Student: Se presenta el intervalo de confianza del parámetro diferencia poblacional de dos medias y los resultados del contraste de hipótesis de dos medias mediante la prueba t-Student.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 1, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

- El programa asume, por defecto, igualdad de varianzas entre las variables, pero puede ser modificado.

F-Snedecor: Se presenta el intervalo de confianza del parámetro cociente poblacional de dos varianzas y los resultados del contraste de hipótesis de dos varianzas mediante la prueba F-Snedecor.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 1, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

- En el intervalo de confianza, el nivel de confianza se toma del valor alfa (nivel de confianza= 100 - α). El programa asigna, por defecto, el valor de α= 5%, pero también son habituales los valores α= 1% y α= 10%. Alfa debe ser >0 y <100.

Dos Grupos (b|y) → Mann-Whitney (Wilcoxon)

Realiza una comparación entre dos distribuciones de datos mediante la prueba de Mann-Whitney, también conocida como prueba de Wilcoxon para dos muestras independientes. Es una prueba no paramétrica.

Se tienen dos variables, una de ellas cuantitativa no normal u ordinal, considerada como variable respuesta (Rta) y la otra dicotómica, considerada como variable explicativa (Exp). Para establecer si hay diferencias en la variable respuesta con relación a los grupos formados por la variable explicativa se utiliza la prueba U de Mann-Whitney o la prueba W de Wilcoxon. Dichas pruebas son equivalentes entre sí y en ambas el contraste que se realiza es

H_0: Las medianas son iguales

H_1: Las medianas son diferentes (caso bilateral)

H_1: La mediana del grupo 1 es superior / inferior a la mediana del grupo 2 (caso unilateral)

La expresión para el cálculo de la U de Mann-Whitney viene dada por U_{XY} o por U_{YX} donde

$$U_{XY} = \#\{x_{i1} < y_{i2}\}$$

$$U_{YX} = \#\{y_{i2} < x_{i1}\}$$

siendo x_{i1} los valores de la variable Rta para el grupo 1 de n_1 individuos, y_{i2} los valores de la variable Rta para el grupo 2 de n_2 individuos.

La notación #{ } representa el número de pares que verifican la condición entre llaves, teniendo en cuenta que cualquier par con $x_{i1} = y_{i2}$ suma 0.5 en el cálculo de U_{XY} y de U_{YX}. Así, U_{XY} sería el número de veces que una observación "x" (del grupo 1) precede a una observación "y" (del grupo 2). De forma análoga U_{XY}, sería el número de veces que una observación "y" (del grupo 2) precede a una observación "x" (del grupo 1). El número de empates entre valores “x” e “y” se denota por d_j, con j = 1,..,k siendo k el número de valores distintos donde se produce el empate. Se verifica que $U_{XY} + U_{YX} = n_1 n_2$.

Las expresiones para $E[U_{XY}]$, $E[U_{YX}]$, $V[U_{XY}]$ y $V[U_{YX}]$ son

$$E[U_{XY}] = E[U_{YX}] = \frac{1}{2} n_1 n_2$$

$$V[U_{XY}] = V[U_{YX}] = \frac{1}{12} \frac{n_1 n_2}{n(n-1)} \left[n^3 - n - \sum_{j=1}^{k} \left(d_j^3 - d_j \right) \right]$$

donde n_1 es el tamaño muestral del grupo 1, n_2 es el tamaño muestral del grupo 2 y n es el tamaño muestral total.

El estadístico de contraste se calcula como

$$z = \frac{\left| U_{XY} - E[U_{XY}] \right|}{\sqrt{V[U_{XY}]}} = \frac{\left| U_{YX} - E[U_{YX}] \right|}{\sqrt{V[U_{YX}]}}$$

que sigue una distribución normal N(0,1).

La expresión para el cálculo de la W de Wilcoxon viene dada por W_1 o por W_2 con

$$W_1 = \sum_{grupo1} Rangos$$

$$W_2 = \sum_{grupo2} Rangos$$

donde los rangos se calculan en relación a la muestra conjunta y en el caso de empates como promedios de los órdenes de las observaciones empatadas. Se verifica que $W_1 + W_2 = 0.5\, n\, (n+1)$.

Las expresiones para $E[W_1]$, $E[W_2]$, $V[W_1]$ y $V[W_2]$ son

$$E[W_1] = \frac{1}{2} n_1 (n+1)$$

$$E[W_2] = \frac{1}{2} n_2 (n+1)$$

$$V[W_1] = V[W_2] = \frac{1}{12} \frac{n_1 n_2}{n(n-1)} \left[n^3 - n - \sum_{j=1}^{k} \left(d_j^3 - d_j \right) \right]$$

donde n_1 es el tamaño muestral del grupo 1, n_2 es el tamaño muestral del grupo 2 y n es el tamaño muestral total.

El estadístico de contraste se calcula como

$$z = \frac{\left|W_1 - E[W_1]\right|}{\sqrt{V[W_1]}} = \frac{\left|W_2 - E[W_2]\right|}{\sqrt{V[W_2]}}$$

que sigue una distribución normal N(0,1). Este programa proporciona un p-valor asintótico para esta opción.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa u ordinal y la *Variable explicativa* dicotómica que forma los dos grupos. La variable respuesta no puede ser constante en los dos grupos.

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Cajas: Gráfico de dos Cajas de la variable respuesta para los dos valores de la variable explicativa. Ver menú Gráficos.

Mann-Whitney (Wilcoxon): Se presentan los resultados del contraste de hipótesis entre dos distribuciones de datos mediante la prueba de Mann-Whitney (Wilcoxon).

Opciones:

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción ‘no igual’, que puede modificarse por ‘mayor que’ o ‘menor que’ (referido a la igualdad de distribuciones que contrasta la hipótesis nula).

Comprobar si la distribución de la variable FC2FC1 es diferente para fumadores y no fumadores.

Resultados descriptivos de la prueba Mann-Whitney (Wilcoxon).

```
Estadísticos para la variable FC2FC1 por FUMADOR
================================================================
-----------------------------------
 Grupos              1          2
-----------------------------------
 N                   16         24
 Media               64.3125    60.0417
 Mediana             67.0000    59.5000
```

```
 Desviación Típica   10.9527    10.8847
 Mínimo              42.0000    42.0000
 Máximo              82.0000    82.0000
 Cuartil Inferior    56.0000    52.0000
 Cuartil Superior    72.0000    68.0000
-----------------------------------------
```

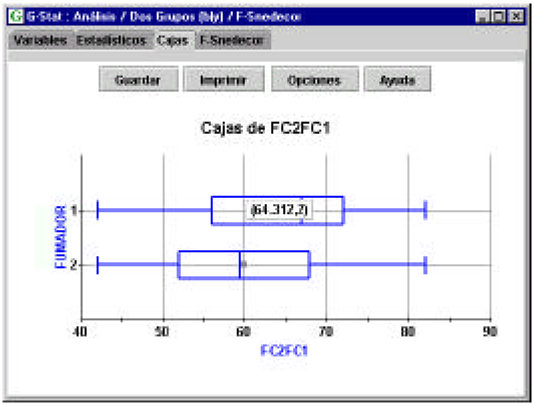


Gráfico de cajas de la opción Mann-Whitney (Wilcoxon)

Resultados de la prueba Mann-Whitney (Wilcoxon).

```
Mann-Whitney (Wilcoxon) de FC2FC1 por FUMADOR
======================================================================

Variable Respuesta:   FC2FC1
Variable Explicativa: FUMADOR

 Grupo                    2          1
--------------------------------------------
 Tamaños Muestrales      24         16
 Medianas:               59.50      67.00
 Rangos Medios:          18.6042    23.3438
--------------------------------------------

Estadístico de Mann-Whitney
---------------------------
Hipótesis Nula:        igualdad de distribuciones
Hipótesis Alternativa: distribución 2 no igual distribución 1

Uxy = 237.5000;  E[Uxy] = 192.0000;  V[Uxy] = 1308.8000
```

```
Uyx = 146.5000;  E[Uyx] = 192.0000;  V[Uyx] = 1308.8000

Estadístico de contraste de U:   -1.2577
p-valor de U:                     0.2085

Estadístico W de Wilcoxon
-------------------------
Hipótesis Nula:        igualdad de distribuciones
Hipótesis Alternativa: distribución 2 no igual distribución 1

W1 = 446.5000;  E[W1] = 492.0000;  V[W1] = 1308.8000
W2 = 373.5000;  E[W2] = 328.0000;  V[W2] = 1308.8000

Estadístico de contraste de W:   -1.2577
p-valor de W:                     0.2085
```

Dos Grupos (b|y) → Wilcoxon. Pareados

Realiza un contraste de hipótesis de la mediana de la distribución de la variable diferencia mediante la prueba de Wilcoxon para datos pareados. Es una prueba no paramétrica.

La prueba de Wilcoxon para datos pareados (o prueba de Rangos signados para datos pareados) calcula en primer lugar una nueva variable Vardif que es igual a la primera variable Var1 menos la segunda variable Var2. Para realizar el contraste

$$H_0: \text{med(Vardif)} = \text{med}_0$$

$$H_1: \text{med(Vardif)} \neq \text{med}_0$$

se aplica a esta nueva variable la prueba de los Rangos signados para una muestra.

Se tiene, por tanto , que la prueba de Wilcoxon para datos pareados se basa en contar, para la variable diferencia, los valores que están por encima y por debajo del valor supuesto para la mediana, teniendo en cuenta las magnitudes además del signo.

El estadístico de contraste sigue una distribución Normal (0,1). Este programa proporciona un p-valor asintótico para esta opción.

Manejo del programa

Variables: Se identifica la *1ª Variable pareada* y la *2ª Variable pareada*, ambas cuantitativas u ordinales y expresadas en las mismas unidades. El programa operará internamente con la diferencia de las dos variables. La variable diferencia no puede ser constante con el valor de la mediana de contraste igual a la mediana de la muestra.

Estadísticos: Estadísticos de la diferencia de las dos variables pareadas. Se detallan en Cuantitativa (y).

Cajas: Gráfico de Cajas para la diferencia de las dos variables pareadas. Ver menú Gráficos.

Wilcoxon. Pareados: Se presentan los resultados del contraste de hipótesis de la mediana de la distribución de la variable diferencia mediante la prueba de Wilcoxon para datos pareados.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

Realizar la prueba de Wílcoxon para observar si la mediana de la diferencia de FC2 menos FC1 es superior a 60.

Resultados de la prueba de Wilcoxon para datos pareados.

```
Wilcoxon para la Mediana Poblacional de FC2 menos FC1
=======================================================================

Tamaño Muestral:   40
Mediana:           61.0000

Wilcoxon. Pareados
------------------

Hipótesis Nula:              mediana = 60.0000
Hipótesis Alternativa:       mayor que
```

```
Rango medio de valores por debajo del valor a contrastar de la mediana:
20.8000
Rango medio de valores por encima del valor a contrastar de la mediana:
16.8824

Estadístico de contraste:     -0.9741 (con corrección de continuidad)
p-valor:                      0.1650
```

Dos Grupos (b|y) → Signos. Pareados

Realiza un contraste de hipótesis de la mediana de la distribución de la variable diferencia mediante la prueba de los Signos para datos pareados. Es una prueba no paramétrica.

La prueba de los Signos para datos pareados calcula en primer lugar una nueva variable Var_dif que es igual a la primera variable Var1 menos la segunda variable Var2. Para realizar el contraste:

$$H_0\text{: med(Var\_dif)} = \text{med}_0$$

$$H_1\text{: med(Var\_dif)} \neq \text{med}_0$$

se aplica a esta nueva variable la prueba de los Signos para una muestra. En el caso de datos dicotómicos pareados, la prueba de los Signos y la de McNemar son equivalentes.

Se tiene, por tanto , que la prueba de Signos para datos pareados se basa en contar, para la variable diferencia, los valores que están por encima y por debajo del valor supuesto para la mediana, sin tener en cuenta las magnitudes, sólo el signo.

El estadístico de contraste sigue una distribución Normal (0,1). Este programa proporciona un p-valor asintótico para esta opción.

Manejo del programa

Variables: Se identifica la *1ª Variable pareada* y la *2ª Variable pareada*, ambas cuantitativas u ordinales y expresadas en las mismas unidades. El programa operará internamente con la diferencia de las dos variables. La variable diferencia no puede ser constante con el valor de la mediana de contraste igual a la mediana de la muestra.

Estadísticos: Estadísticos de la diferencia de las dos variables pareadas. Se detallan en Cuantitativa (y).

Cajas: Gráfico de Cajas para la diferencia de las dos variables pareadas. Ver menú Gráficos.

Signos. Pareados: Se presentan los resultados del contraste de hipótesis de la mediana de la distribución de la variable diferencia mediante la prueba de Signos para datos pareados.

Opciones:

- El valor que se quiere contrastar en la Hipótesis Nula. El programa asigna, por defecto, el valor 0, que es el más habitual, pero puede modificarse.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción *'no igual'*, que puede modificarse por *'mayor que'* o *'menor que'* (referido al valor que contrasta la hipótesis nula).

Dos Grupos (b|y cens)

Abre un submenú con una técnica estadística para una variable explicativa dicotómica o binaria (b) que intenta explicar una variable respuesta cuantitativa con datos censurados por la derecha (y cens). Se incluye la prueba Log-Rank.

Dos Grupos (b|y cens) → Log-Rank

Realiza una comparación entre dos curvas de supervivencia, donde la supervivencia representa la probabilidad de sobrevivir a un tiempo dado, mediante la prueba Log-Rank.

Para realizar la prueba Log-Rank, se ordenan de forma creciente los k distintos valores exactos (no censurados) de la variable respuesta ("tiempo hasta"), denotando estos tiempos por

$$t_1 < t_2 < \ldots < t_k$$

En cada uno de estos tiempos se construye una tabla 2x2 con

	Grupo 1	Grupo 2	Total
Muertes	d_{1j}	d_{2j}	d_j
Supervivientes	n_{1j}-d_{1j}	n_{2j}-d_{2j}	n_j-d_j
En riesgo	n_{1j}	n_{2j}	n_j

siendo d_{1j} el número de individuos del grupo 1 con dato exacto igual a t_j, d_{2j} el número de individuos del grupo 2 con dato exacto igual a t_j, $d_j=d_{1j}+d_{2j}$. Los individuos a riesgo n_{1j} del grupo 1 son aquellos que tienen dato exacto mayor o igual que t_j, análogamente con los individuos en riesgo del grupo 2, $n_j=n_{1j}+n_{2j}$.

A partir de las k tablas 2x2 anteriores se construye el estadístico de contraste z

$$z=\frac{\sum_{j=1}^{k}\left(d_{1j}-n_{1j}\frac{d_j}{n_j}\right)}{\sqrt{\sum_{j=1}^{k}\frac{n_{1j}n_{2j}d_j\left(n_j-d_j\right)}{n_j^2\left(n_j-1\right)}}}$$

que sigue una distribución Normal o el estadístico Chi-2

$$\chi^2=\frac{\left(\sum_{j=1}^{k}\left(d_{1j}-n_{1j}\frac{d_j}{n_j}\right)\right)^2}{\sum_{j=1}^{k}\frac{n_{1j}n_{2j}d_j\left(n_j-d_j\right)}{n_j^2\left(n_j-1\right)}}$$

que sigue una distribución Chi-Cuadrado con 1 grado de libertad.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa junto con la variable dicotómica que contiene el código de dato censurado y la *Variable explicativa* dicotómica que forma los grupos. El código asociado al dato censurado de la variable censura puede cambiarse en las opciones de las pantallas de resultados.

Estadísticos: Estadísticos de la variable respuesta y de la variable explicativa estratificados por la variable censura. Se detallan en Cuantitativa (y). Los estadísticos a calcular se definen en las opciones.

Kaplan-Meier Tabla: Datos de la curva de supervivencia de todos los casos y , alternativamente, las de los grupos formados por la variable explicativa.

Opciones:

- El código asociado a los datos censurados.

- Estratificación por la variable explicativa.

Kaplan-Meier Gráfico: Gráfico con la curva de supervivencia de todos los casos y , alternativamente, las de los grupos formados por la variable explicativa.

Opciones:

- El código asociado a los datos censurados.

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento de los ejes X e Y.

- Estratificación por la variable explicativa.

Log-Rank: Se presentan los resultados del contraste de hipótesis entre dos curvas de supervivencia mediante la prueba Log-Rank.

Opciones:

- El código asociado a los datos censurados.

- El tipo de aproximación deseada en la Hipótesis Alternativa. El programa asigna, por defecto, la opción ‘no igual’, que puede modificarse por ‘mayor que’ o ‘menor que’ (referido a la igualdad de curvas de supervivencia que contrasta la hipótesis nula).

x|y

Abre un submenú con diferentes técnicas estadísticas para una variable explicativa cuantitativa (x) que intenta explicar una variable cuantitativa (y) y técnicas de asociación para dos variables cuantitativas. Son, por tanto, técnicas de regresión y correlación. Se incluyen las siguientes técnicas: Regresión Lineal Simple, Modelos Transformados y Regresión Polinómica.

x|y → Regresión Lineal Simple

Realiza el análisis de regresión lineal simple, proporcionando, fundamentalmente, los coeficientes de la regresión y su significación. De forma adicional se calculan los coeficientes de correlación de Pearson, Spearman e intraclase.

El cálculo de los coeficientes b_0 y b_1 se realiza mediante el método de los mínimos cuadrados. La significación estadística de los coeficientes del modelo se calcula a partir del coeficiente y de su error estándar, de forma que para b_0

$$t(b_0) = \frac{b_0}{EE(b_0)}$$

sigue una distribución t-Student con n - 2 grados de libertad y para b_1

$$t(b_1) = \frac{b_1}{EE(b_1)}$$

sigue una distribución t-Student con n – 2 grados de libertad.

Para realizar el contraste

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

a través del coeficiente de correlación de Pearson, se construye el siguiente estadístico de contraste

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

que sigue una distribución t-Student con n – 2 grados de libertad. La significación del coeficiente de correlación de Pearson coincide con la significación de la pendiente de la ecuación de la recta de regresión.

Para realizar el contraste

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

a través del coeficiente de correlación de Spearman se construye el siguiente estadístico de contraste

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

que sigue una distribución t-Student con n - 2 grados de libertad.

Adicionalmente, el programa muestra el coeficiente de correlación intraclase CCI y su significación. Dicho coeficiente se calcula creando una disposición de datos intermedia de la forma

1	1	y_1
1	2	x_1
2	1	y_2
2	2	x_2
...		
...		
n	1	y_n
n	2	x_n

y utilizando la técnica Anova un factor con bloques (ver más adelante en el manual en la opción Anova / Anova un factor con bloques), que proporciona el CMB (cuadrado medio entre bloques), el CMG (cuadrado medio entre grupos), el CMR (cuadrado medio residual) y el valor F de los bloques. A partir de estos valores se calcula

$$CCI = \frac{n(CMB - CMR)}{nCMB + 2CMG + (2n - n - 2)CMR}$$

La significación del CCI viene dada por la significación del valor F de los bloques.

Cuando la variable que se intenta explicar muestra inestabilidad en la varianza (los valores grandes tienen mayor varianza que los valores pequeños) y se desconoce su distribución, es posible aplicar la transformación de Box-Cox, método diseñado para hallar la mejor transformación de la variable respuesta. Las transformaciones más frecuentes son:

Modelo	**Expresión**
Exponencial	$y = \exp(b_0 + b_1 x)$
Recíproco-y	$y = \frac{1}{b_0 + b_1 x}$
Recíproco-x	$y = b_0 + b_1 \frac{1}{x}$

Recíproca-doble	$\frac{1}{y} = b_0 + b_1 \frac{1}{x}$
Logaritmo-x	$y = b_0 + b_1 Ln(x)$
Multiplicativo	$y = b_0 x^{b1}$
Raíz cuadrada-x	$y = b_0 + b_1 \sqrt{x}$
Raíz cuadrada-y	$\sqrt{y} = b_0 + b_1 x$
Curva en S	$y = \exp\left(b_0 + b_1 \frac{1}{x}\right)$

Todas ellas tienen una expresión analizable mediante un modelo de regresión simple utilizando las variables transformadas. De todas las posibles transformaciones es habitual escoger como la más adecuada, aquella que presenta un coeficiente de determinación R^2 mayor, o bien en función de la representación gráfica de las variables.

Puede que, a pesar de hacer las transformaciones anteriores no se consiga linealidad en la nube de puntos. En estos casos, se puede recurrir a modelos de regresión polinómica que son un caso particular de la regresión multivariante, pero donde sólo se utiliza la variable explicativa original y potencias de ésta. En los modelos de regresión polinómica, de forma análoga a la regresión lineal simple, existen coeficientes del modelo, significaciones de éstos y coeficiente R^2 del modelo.

Recta de ajuste

En el gráfico se muestra la recta de regresión estimada por mínimos cuadrados. Además se muestra el huso de predicciones que consta de unas curvas interiores al 95% de predicción para valores medios (color rojo) y de unas curvas exteriores al 95% de predicción de valores individuales (color azul).

Las líneas de predicción para valores medios (rojas) del gráfico se calculan como

$$b_0 + b_1 x_i \pm t_{1-0.05/2,n-2} \cdot s_e \sqrt{\frac{1}{n} + \frac{\left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}$$

Las líneas de predicción de valores individuales (azules) del gráfico se calculan como

$$b_0 + b_1 x_i \pm t_{1-0.05/2,n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

Los valores x_i no se limitan a los valores observados en la muestra, sino a cualquier valor comprendido en el rango experimental.

Anova en la regresión

La tabla Anova del modelo permite establecer la significación global del modelo. Para su cálculo intervienen distintos estadísticos: suma de cuadrados del modelo (SCM), suma de cuadrados total (SCT), suma de cuadrados residual (SCR), grados de libertad del modelo (GLM), grados de libertad total (GLT), grados de libertad residual (GLR), cuadrado medio del modelo (CMM) y cuadrado medio residual (CMR), donde

$$SCM = \frac{\left[\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}\right]^2}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}$$

$$SCT = \left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right]$$

$$SCR = SCT - SCM$$

$$GLM = 1\,,\ GLT = n - 1\,,\ GLR = GLT - GLM$$

$$CMM = \frac{SCM}{GLM}\,,\ CMR = \frac{SCR}{GLR}$$

Por último, la significación del modelo viene dada por

$$F = \frac{CMM}{CMR}$$

que sigue una distribución F-Snedecor con GLM = 1 grados de libertad del numerador y GLR = n – 2 grados de libertad del denominador.

Si se calcula el cociente entre SCM y SCT se obtiene el coeficiente de determinación, que si se expresa en tanto por ciento representa el porcentaje de información que explica el modelo. El coeficiente de determinación también se puede calcular como el cuadrado del coeficiente de correlación de Pearson.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* cuantitativa.

Estadísticos: Estadísticos univariantes de la variable respuesta y de la variable explicativa. Se detallan en Cuantitativa (y).

Recta de Ajuste: Gráfico con la recta de regresión estimada por mínimos cuadrados. Además, se muestra el huso de predicciones que consta de unas curvas interiores al 95% de predicción para valores medios (color rojo) y de unas curvas exteriores al 95% de predicción de valores individuales (color azul). Posicionando el cursor sobre cualquier punto se obtienen las coordenadas.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento de los ejes X e Y.

Modelo: Resultados de la regresión lineal simple. Coeficientes de la regresión, errores estándar, significación de los coeficientes, desviación típica de los residuos. Correlación de Pearson. Correlación de Spearman con su significación. Coeficiente de correlación intraclase y significación.

Opciones:

- El tipo de modelo que se desea ajustar: Lineal, Exponencial, Recíproco-y, Recíproco-x, Recíproca Doble, Logaritmo-x, Multiplicativo, Raíz Cuadrada-x, Raíz Cuadrada-y, Curva en S.

Anova: Significación de la regresión lineal mediante el análisis de la varianza. La significación del modelo coincide con la pendiente o coeficiente de regresión b1.

Establecer si la variable FC1 está relacionada o explica significativamente la variable FC2.

Estadísticos de la opción regresión lineal simple.

```
Regresión Lineal Simple. Estadísticos
====================================================================

Variable Respuesta:           FC2
Variable Explicativa:         FC1
Número de Casos:              40

--------------------------------------------------
 Variable                    FC1        FC2
--------------------------------------------------
 N                           40.0       40.0
 Media                       75.9500    137.7000
 Mediana                     74.0000    137.0000
 Moda                        68.0000    126.0000
 Media Geométrica            75.3846    137.0738
 Varianza                    90.9718    176.3692
 Desviación Típica           9.5379     13.2804
 E.E. de la Media (*)        1.5081     2.0998
 Mínimo                      62.0000    112.0000
 Máximo                      96.0000    165.0000
 Rango                       34.0000    53.0000
 Cuartil Inferior            68.0000    127.0000
 Cuartil Superior            83.0000    147.0000
 Rango Intercuartílico       15.0000    20.0000
 Asimetría                   0.5493     0.0850
 Asimetría  Estandarizada    1.4184     0.2196
 Curtosis                    -0.7245    -0.6570
 Curtosis Estandarizada      -0.9353    -0.8482
 Coeficiente de Variación    12.5581    9.6445
--------------------------------------------------

(*) Usar con propósito de estimación para el I.C. de la media
```

Resultados de la regresión lineal simple.

```
Modelo de FC2 con FC1
======================================================================

Número de Casos: 40

Modelo: Lineal
----------------------------------------------------------------
Ecuación: FC2 = 76.4031 + 0.8071 * FC1
----------------------------------------------------------------
              Coef.        E.E.         t-valor      p-valor
```

```
-----------------------------------------------------------------
 Ordenada     76.4031      14.0864      5.4239       0.0001
 Pendiente    0.8071       0.1841       4.3848       0.0001
-----------------------------------------------------------------
r de Pearson (coeficiente de correlación)  0.5796
r cuadrado (coeficiente de determinación)  33.60%
Desviación Típica de los Residuos          10.9634
Rho de Spearman 0.5567   t-valor 4.1311   p-valor 0.0002
```

Significación mediante la prueba Anova.

```
Regresión Lineal Simple. Análisis de la Varianza
=======================================================================

Variable Respuesta:    FC2
Variable Explicativa:  FC1
Número de Casos:       40

Anova
Modelo: Lineal
                   Suma de            Cuadrado
 Variabilidad      Cuadrados    G.L.  Medio      F-valor      p-valor
-------------------------------------------------------------------------
 Modelo            2310.9613    1     2310.9613  19.2266      0.0009E-1
 Residual          4567.4387    38    120.1958
-------------------------------------------------------------------------
   Falta de ajuste 2211.3054    15    147.4204   1.4391       0.2101
   Error           2356.1333    23    102.4406
-------------------------------------------------------------------------
 Total                   6878.4000     39
```

x|y → Modelos Transformados

Calcula el coeficiente de determinación, R cuadrado, para distintos modelos de regresión simple.

El coeficiente de determinación se calcula como el cociente entre la suma de cuadrados explicada y la suma de cuadrados total, por lo que cuantifica la proporción de variabilidad de la variable respuesta que es explicada por la variable explicativa, o dicho de otra forma, el porcentaje de información compartida. El coeficiente de determinación se encuentra entre 0 y 1 (ó entre 0 y 100 si se expresa en tanto por ciento). El mejor modelo es el de mayor R cuadrado.

Los modelos disponibles se encuentran descritos en Análisis / x|y / Regresión Lineal Simple.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* cuantitativa.

Modelos: Muestra los resultados del coeficiente de determinación para los diferentes modelos construidos. Los modelos son: Lineal, Exponencial, Recíproco-y, Recíproco-x, Recíproca Doble, Logaritmo-x, Multiplicativo, Raíz Cuadrada-x, Raíz Cuadrada-y, Curva en S.

Establecer los coeficientes de determinación de todos los posibles modelos de regresión entre las variables FC2 y FC1.

Coeficientes de determinación de modelos transformados.

```
Comparación de Modelos Transformados de Y = FC1 con X = FC2
======================================================================
Número de Casos: 40

 Modelo             r-cuadrado(%)
----------------------------------
 Lineal             33.5974
 Exponencial        33.3378
 Recíproco-Y        32.8187
 Recíproco-X        31.7372
 Recíproca Doble    31.1088
 Multiplicativo     32.5389
 Raiz Cuadrada-X    33.1994
 Raiz Cuadrada-Y    33.5000
 Curva en S         31.5561
----------------------------------
```

x|y → Regresión Polinómica

Realiza un análisis de regresión polinómica para modelos bivariantes cuadráticos y cúbicos.

En la regresión polinómica se contempla como posibles variables explicativas a la variable x y a potencias de ésta. Es, por tanto, un caso particular de la Regresión Lineal Múltiple. Ver fórmulas en Multivariante / Regresión Múltiple. Se contemplan ajustes polinómicos de grado 2 ó 3 según se especifique. Si la significación del coeficiente de mayor orden es <0.05 se tiene que el modelo con dicho término es mejor que sin él. El coeficiente de determinación R2 refleja el porcentaje de variabilidad explicada por el modelo. El estadístico de Durbin-Watson se usa para estudiar la autocorrelación en los residuos.

La tabla Anova del modelo permite establecer la significación global del modelo. Para su cálculo intervienen distintos estadísticos: suma de cuadrados del modelo (SCM), suma de cuadrados total (SCT), suma de cuadrados residual (SCR), grados de libertad del modelo (GLM), grados de libertad total (GLT), grados de libertad residual (GLR), cuadrado medio del modelo (CMM) y cuadrado medio residual (CMR). Ver fórmulas en Multivariante / Regresión Múltiple.

La significación del modelo viene dada por F que se calcula como el cociente entre CMM y CMR,

que sigue una distribución F-Snedecor con GLM = p (con p el orden del polinomio) grados de libertad del numerador y GLR = n – 1 – p grados de libertad del denominador.

Si se calcula el cociente entre SCM y SCT se obtiene el coeficiente de determinación, que si se expresa en tanto por ciento, representa el porcentaje de información que explica el modelo.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* cuantitativa.

Estadísticos: Estadísticos univariantes de la variable respuesta y de la variable explicativa. Se detallan en Cuantitativa (y).

Modelo: Resultados de la regresión polinómica. Coeficientes de la regresión, errores estándar, significación de los coeficientes, desviación típica de los residuos.

Opciones: Orden del polinomio: 2 ó 3.

Anova: Resultados de la significación de los modelos mediante el análisis de la varianza.

Orden del Polinomio: Se facilita una descripción de distintos modelos polinómicos con distinto grado, ya que pudiera ocurrir que un modelo polinómico de mayor grado no mejorara estadísticamente un modelo polinómico con grado inferior.

Opciones: Orden del polinomio: 2 ó 3.

Comparar los modelos polinómicos de regresión que pueden conformarse con la variable Edad como variable explicativa y la variable FC2FC1 como variable respuesta.

Resultados de la regresión polinómica de orden 2 en la pestaña Modelo.

```
Regresión Polinómica. Modelo
======================================================================

Variable Respuesta:     FC2FC1
Variable Explicativa:   EDAD
Número de Casos:         40

Modelo Polinomial de orden 2
--------------------------------------------------------
            Coef.       E.E.        t-valor    p-valor
--------------------------------------------------------
 Ordenada   289.7652    118.0072    2.4555     0.0189
 EDAD       -13.1653    10.4859     -1.2555    0.2172
 EDAD^2     0.1330      0.2324      0.5723     0.5706
--------------------------------------------------------

r cuadrado (coeficiente de determinación)    88.3239
r cuadrado (ajustado)                        87.6928%
Desviación Típica de los Residuos             3.8510
Error Absoluto Medio                          2.8737
Durbin-Watson                                 1.5419
```

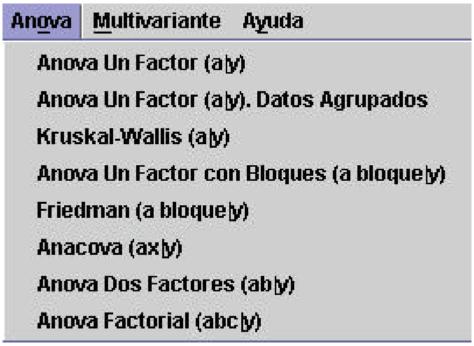
Comparación de modelos polinómicos con la pestaña Orden de polinomio.

```
Regresión Polinómica. Orden del Polinomio
======================================================================

Variable Respuesta:         FC2FC1
Variable Explicativa:       EDAD
Número de Casos:            40

ANOVA para Modelos con distinto orden de polinomio ajustado.
----------------------------------------------------------------------
```

Fuente	Suma de Cuadrados	G.L.	Cuadrado Medio	F-valor	p-valor	r cuadrado
EDAD	4145.9258	1	4145.9258	283.8489	0.0001E-14	88.2206
EDAD^2	4.8565	1	4.8565	0.3325	0.5678	88.3239
EDAD^3	22.8982	1	22.8982	1.5677	0.2186	88.8112
Modelo	4173.6805					

El modelo no mejora significativamente con la utilización de polinomios de grado 2 ó 3.

Menú Anova



Activar la opción **Anova** del menú principal o mediante Alt+O. Este menú contiene las opciones necesarias para realizar las diferentes pruebas Anova.

Anova Un Factor (a|y)

Realiza la prueba Anova un factor. Asimismo, facilita, entre otras técnicas asociadas a ella, las comparaciones múltiples a posteriori.

El Análisis de la Varianza (Anova: "Analysis of Variance") permite comparar las medias de r grupos, siendo r mayor o igual a 2. El modelo Anova presupone que las varianzas de los grupos son iguales y que los residuos o errores son aleatorios, independientes e idénticamente distribuidos siguiendo una ley normal con media 0 y desviación constante. La hipótesis nula de la prueba Anova de un factor es:

H_0: Las medias de los k grupos son todas iguales

H_1: Al menos una de las medias es diferente

Esta prueba se basa en la comparación de las sumas de cuadrados medias, debidas a la variabilidad entre grupos y la debida a la variabilidad intra grupos (dentro de los grupos). Ambas sumas son estimaciones independientes de la variabilidad global, de manera que, si el cociente entre la primera y la segunda es grande, se tendrá mayor probabilidad de rechazar la hipótesis nula. Este cociente sigue una distribución F con r - 1 y n - r grados de libertad. La hipótesis nula de igualdad de medias se rechaza en el caso en el que p-valor<0.05, en caso contrario no hay evidencia suficiente para poder rechazarla. En el caso de que se rechace la hipótesis nula de igualdad de medias se puede determinar mediante comparaciones múltiples a posteriori, de qué grupo o grupos provienen esas diferencias.

Las sumas de cuadrados son un paso previo para el cálculo del Anova. La suma de cuadrados entre grupos SCE, la suma de cuadrados dentro de grupos SDE y la suma de cuadrados total SCT se calculan del siguiente modo:

$$SCE = \sum_{j=1}^{r} n_j \left(\bar{x}_{.j} - \bar{x}_{..}\right)^2$$

$$SCD = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_{.j}\right)^2 = \sum_{j=1}^{r}\sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{r} n_j \bar{x}_{.j}^2$$

$$SCT = \sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_{..}\right)^2$$

donde se denota por r al número de grupos, por n_j el número de individuos en cada grupo j = 1,...,r, $\bar{x}_{.j}$ la media de cada grupo y $\bar{x}_{..}$ la media global.

Utilizando la siguiente igualdad que permite expresar las desviaciones entre los datos observados x_{ij} y la media total (“grand mean”) $\bar{x}_{..}$ como suma de las desviaciones de la media del grupo $\bar{x}_{.j}$ y la media total más las desviaciones entre los datos observados y la media del grupo, de forma que

$$x_{ij} - \bar{x}_{..} = \left(\bar{x}_{.j} - \bar{x}_{..}\right) + \left(x_{ij} - \bar{x}_{.j}\right)$$

se puede demostrar que SCT = SCE + SCD y, por tanto, la variabilidad de los datos (dada por SCT) se expresa como la suma de la variabilidad explicada debida a los grupos (a las medias) dada por SCE más la variabilidad dentro de los grupos (variabilidad residual o variabilidad no explicada) dada por SCD.

Los grados de libertad entre grupos GLE, dentro de los grupos GLD y total GLT son

$$GLE = r - 1 \text{ , } GLD = n - r \text{ , } GLT = n - 1$$

El cuadrado medio entre grupos CME y el cuadrado medio dentro de grupos son

$$CME = \frac{SCE}{GLE}, \quad CMD = \frac{SCD}{GLD}$$

El estadístico de contraste para realizar la prueba Anova se construye con

$$F = \frac{CME}{CMD}$$

que se distribuye según una F-Snedecor con GLE grados de libertad del numerador y GLD grados de libertad del denominador.

Una medida relativa de la variabilidad explicada por los grupos es el cociente

$$R^2 = \frac{SCE}{SCT}$$

que se denomina coeficiente de determinación; este coeficiente estará entre cero y uno. Queda claro que cuanto más próximo esté de 1, más variabilidad explica el modelo, y, por tanto, menos variabilidad no explicada o residual.

La información anterior se suele disponer en forma de tabla:

	Suma de Cuadrados	G.L.	Cuadrado Medio	F-valor	p-valor
Entre Grupos	SCE	GLE	CME	F	p
Dentro Grupos	SCD	GLD	CMD		
Total	SCT	GLT			

Medias e IC

Se muestran para cada uno de los grupos, las medias de la variable cuantitativa, junto con su error estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña "Gráfico de Medias".

Existen las siguientes opciones:

- Ninguno: no se muestra ningún intervalo.
- Errores Estándar (agrupado): intervalos dados por la media +/- error estándar basado en la varianza conjunta (pooled). En esta opción el error estándar para la media de cada grupo i y los límites del intervalo vienen dados por

$$\sqrt{\frac{CMD}{n_j}}\ ,\quad \bar{x}_j \pm \sqrt{\frac{CMD}{n_j}}$$

- Errores Estándar (individual): intervalos dados por la media +/- error estándar basado en cada una de las varianzas individuales. En esta opción el error estándar para la media de cada grupo j y los límites del intervalo vienen dados por

$$\sqrt{\frac{s_j^2}{n_j}}\ ,\ \ \bar{x}_j \pm \sqrt{\frac{s_j^2}{n_j}}$$

- Intervalos de Confianza (agrupado): intervalos de confianza para cada media basados en la varianza conjunta. En esta opción el error estándar para la media de cada grupo j y los límites del intervalo vienen dados por

$$\sqrt{\frac{CMD}{n_j}}\ ,\ \ \bar{x}_j \pm t_{1-\alpha/2;\,n-r}\sqrt{\frac{CMD}{n_j}}$$

- Intervalos de Confianza (individual): intervalos de confianza para cada media basados en la varianza individual. En esta opción el error estándar para la media de cada grupo j y los límites del intervalo vienen dados por

$$\sqrt{\frac{s_j^2}{n_j}}\ ,\ \ \bar{x}_j \pm t_{1-\alpha/2;\,n_j-1}\sqrt{\frac{s_j^2}{n_j}}$$

- Intervalos LSD: intervalos basados en el método de Fisher de mínima diferencia significativa (least significant difference = LSD), donde se comparan cualquier par de medias con una confianza prefijada. En esta opción el error estándar para la media de cada grupo j y los límites del intervalo vienen dados por

$$\sqrt{\frac{CMD}{n_j}}\ ,\ \ \bar{x}_j \pm \sqrt{F_{1-\alpha;\,1;\,n-r}}\sqrt{\frac{CMD}{n_j}}$$

- Intervalos HSD de Tukey: intervalos basados en el método de Tukey de diferencia “honradamente” significativa (honestly significant difference = HSD), donde se comparan cada par de medias con una confianza prefijada. Si se denota por q a la función “rango estudentizado inversa”, en esta opción el error estándar para la media de cada grupo j y los límites del intervalo vienen dados por

$$\sqrt{\frac{CMD}{n_j}}\ ,\ \ \bar{x}_j \pm q_{1-\alpha;\,n-1;\,r}\sqrt{\frac{1}{2}}\sqrt{\frac{CMD}{n_j}}$$

- Intervalos Scheffé: intervalos basados en el método de Scheffé, donde se calculan todos los contrastes con al menos una confianza prefijada. En esta opción el error estándar para la media de cada grupo j y los límites del intervalo vienen dados por

$$\sqrt{\frac{CMD}{n_j}}\ ,\ \ \bar{x}_j \pm \sqrt{F_{1-\alpha;\,r;\,n-r}}\sqrt{r}\sqrt{\frac{CMD}{n_j}}$$

- Intervalos de Bonferroni: intervalos basados en el método de Bonferroni, donde se calculan los contrastes seleccionados con al menos una confianza prefijada. En esta opción el error estándar para la media de cada grupo j y los límites del intervalo vienen dados por

$$\sqrt{\frac{CMD}{n_j}}\ ,\quad \bar{x}_j \pm \sqrt{F_{1-\alpha_b\,;\,1;\,n-r}}\sqrt{\frac{CMD}{n_j}}\ ,\quad \alpha_b = \frac{\alpha}{\frac{r(r-1)}{2}}$$

Comparaciones Múltiples

Se presentan pruebas a posteriori para determinar de qué grupo provienen las diferencias detectadas en el ANOVA. Mediante un asterisco se señalan los grupos que son diferentes y mediante un aspa se agrupan los grupos homogéneos o semejantes.

En la pantalla de opciones se selecciona el método que se desee para la formación de grupos homogéneos de las medias. Existen las siguientes posibilidades:

- LSD: basado en el método de Fisher de mínima diferencia significativa (least significant difference = LSD), donde se comparan cualquier par de medias con una confianza prefijada. En esta opción el límite viene dado por

$$\sqrt{2\cdot F_{1-\alpha;\,1;\,n-r}}\sqrt{\frac{CMD}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

- HSD de Tukey: basado en el método de Tukey de diferencia “honradamente” significativa (honestly significant difference = HSD), donde se comparan cada par de medias con una confianza prefijada. En esta opción el límite viene dado por

$$q_{1-\alpha;\,n-r;\,r}\sqrt{\frac{CMD}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

- Scheffé: basado en el método de Scheffé, donde se calculan todos los contrastes con al menos una confianza prefijada. En esta opción el límite viene dado por

$$\sqrt{F_{1-\alpha;\,r-1;\,n-r}}\sqrt{2\cdot(r-1)}\sqrt{\frac{CMD}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

- Bonferroni: basado en el método de Bonferroni, donde se calculan los contrastes seleccionados con al menos una confianza prefijada. En esta opción el límite viene dado por

$$\sqrt{2 \cdot F_{1-\alpha_b;\,1;\,n-r}}\sqrt{\frac{CMD}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

$$\alpha_b = \frac{\alpha}{\frac{r(r-1)}{2}}$$

Homocedasticidad

El ANOVA requiere que la variabilidad dentro de los grupos sea homogénea (hipótesis de homocedasticidad). La hipótesis nula sería la homogeneidad de los grupos (homocedasticidad) y la alternativa la no homogeneidad (heterocedasticidad). Se presentan tres pruebas que contrastan este aspecto: C de Cochran, Bartlett y Levene. En el caso de que se obtenga una $p<0.05$ se tendría que no se verificaría este supuesto, con lo que habría que intentar una transformación (p.ej. la logarítmica) de los datos de la variable cuantitativa y luego aplicar el ANOVA.

El estadístico de Cochran se calcula como

$$\frac{\text{máx}\left\{s_j^2\right\}}{\sum s_j^2}$$

El estadístico de Bartlett se calcula como

$$\frac{\sum (n_j-1) Lns^2 - \sum (n_j-1) Lns_j^2}{1+\frac{1}{3(r-1)}\left(\sum \frac{1}{(n_j-1)} - \frac{1}{\sum (n_j-1)}\right)}$$

con s^2 = CMD.

Para calcular el estadístico de Levene se realiza, en primer lugar, una transformación de la variable respuesta, considerando el valor absoluto de la diferencia entre el valor original y la media del grupo a la que pertenece la observación. Es decir, la nueva variable respuesta es

$$\left|x_{ij} - \overline{x}_{.j}\right|$$

En segundo lugar se realiza un Anova de un factor con esta nueva variable respuesta. El valor F de dicho Anova es el estadístico F de Levene.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y la *Variable explicativa* cualitativa que forma los grupos. La variable respuesta no puede ser constante. La variable explicativa debe tener dos o más grupos.

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Dispersión: Se muestran los datos de los casos para la variable respuesta según las categorías de la variable explicativa o factor. Así, para cada uno de los niveles del factor, que aparecen en el eje X, pueden verse los valores de la variable respuesta observados. Este gráfico permite tener una aproximación visual de cuál es el efecto del factor sobre la variable respuesta respecto de su media y de su dispersión.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Cajas: Gráfico de Cajas de la variable respuesta para los distintos valores de la variable explicativa. Ver menú Gráficos.

Anova: Resultados del Análisis de la Varianza para la comparación de medias de los distintos grupos.

Medias e IC: Se muestran para cada uno de los grupos, las medias de la variable respuesta, junto con sus errores estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña “Gráfico de Medias”.

Opciones:

- Método: Ninguno, Errores Estándar (agrupado), Errores Estándar (individual), Intervalos de Confianza (agrupado), Intervalos de Confianza (individual), Intervalos LSD, Intervalos HSD de Tukey, Intervalos Scheffé e Intervalos Bonferroni

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

Gráfico de Medias: Se muestran para cada uno de los grupos, las medias de la variable cuantitativa, junto con sus intervalos de confianza calculados según distintos métodos.

Opciones:

- Método: Son los dados en la pestaña "Medias e IC".

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Comparaciones Múltiples: Resultados de las pruebas a posteriori para determinar de qué grupo o grupos provienen las diferencias detectadas en el Anova. Este programa se basa en los resultados de las comparaciones dos a dos obtenidas. Mediante un asterisco se señalan los grupos que son diferentes y mediante un aspa se agrupan los grupos homogéneos o semejantes.

Opciones:

- Método: LSD, HSD de Tukey, Scheffé, Bonferroni.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

Homocedasticidad: Resultados del contraste de la homogeneidad de la variabilidad dentro de los grupos mediante las pruebas C de Cochran, Bartlett y Levene.

Residuos vs Predicciones: Se representan los residuales frente a los valores que predice el modelo. Este gráfico sirve para poder detectar falta de homocedasticidad (heterocedasticidad). La banda de residuos debería ser similar en dispersión y simetría a lo largo de todos los predichos.

Opciones:

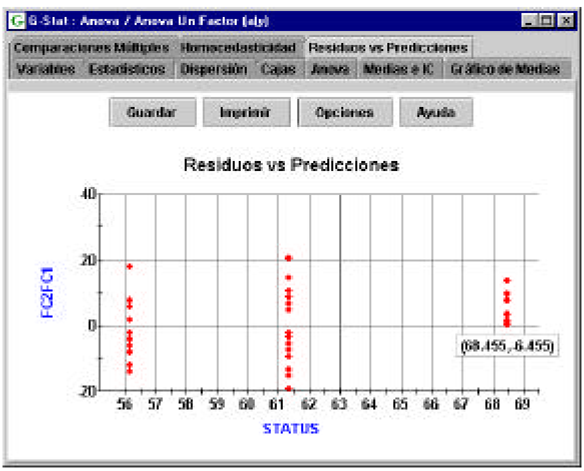
- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Se desea realizar la prueba Anova un factor de la variable FC2FC1 con la variable Status como explicativa.

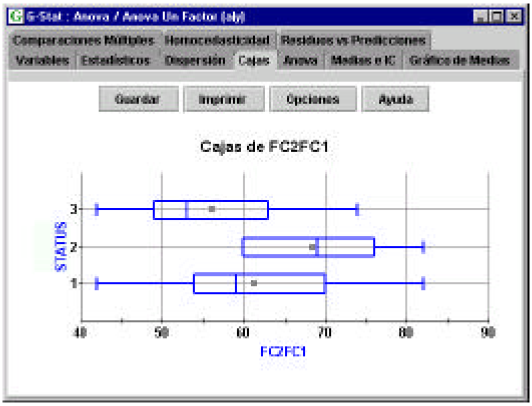
Estadísticos de la opción Anova Un Factor.

Estadísticos para la variable FC2FC1 por STATUS

Grupos	1	2	3
N	17	11	12
Media	61.3529	68.4545	56.1667
Mediana	59.0000	69.0000	53.0000
Desviación Típica	11.1183	7.8913	10.5644
Mínimo	42.0000	60.0000	42.0000
Máximo	82.0000	82.0000	74.0000



Residuos frente a Predicciones de la opción Anova Un Factor.



Cajas de la opción Anova Un Factor.

Resultados de la prueba Anova Un Factor.

Anova Un Factor

==

Variable Respuesta: FC2FC1
Variable Explicativa: STATUS
Número de Casos: 40

	Suma de Cuadrados	G.L.	Cuadrado Medio	F-valor	p-valor
Entre Grupos	871.2237	2	435.6119	4.2102	0.0225
Dentro Grupos	3828.2763	37	103.4669		
Total (corr.)	4699.5000	39			

Medias e IC de la opción Anova Un Factor con los intervalos LSD.

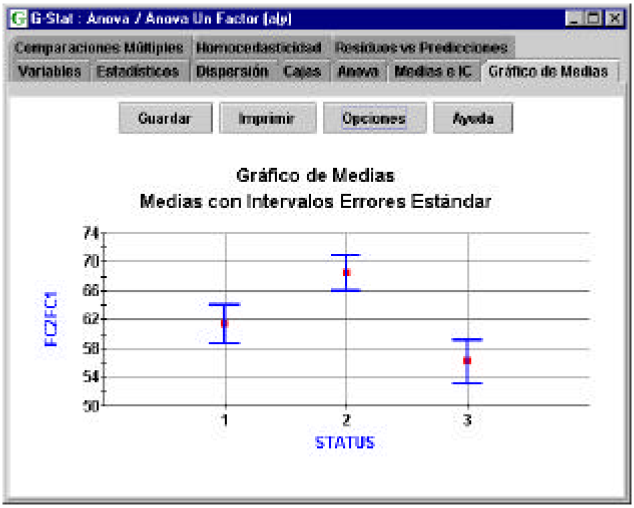
Anova Un Factor

==

Variable Respuesta: FC2FC1
Variable Explicativa: STATUS
Número de Casos: 40

Tabla de Medias con I.C. LSD al 95.0%

STATUS	N	Media	E.E. (agrupado)	Límite Inferior	Límite Superior
1	17	61.3529	2.4670	56.3542	66.3516
2	11	68.4545	3.0669	62.2403	74.6687
3	12	56.1667	2.9364	50.2170	62.1163
Total	40	61.7500			



Gráficos de Medias con Intervalos errores estándar

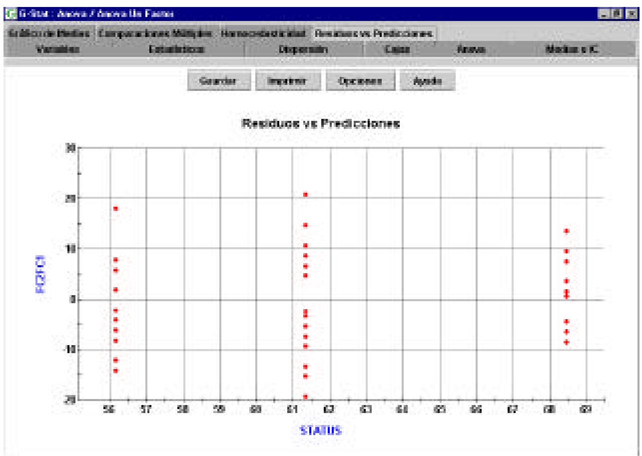
Comparaciones Múltiples LSD de la opción Anova Un Factor.

```
Anova Un Factor. Comparaciones Múltiples
======================================================================
Variable Respuesta:          FC2FC1
Variable Explicativa:        STATUS
Número de Casos:              40

Método: LSD al 95.00%
--------------------------------------
                          Grupos
 STATUS    N    Media      Homogéneos
--------------------------------------
 3         12   56.1667     X
 1         17   61.3529     XX
 2         11   68.4545      X

--------------------------------------
```

```
 Contraste   Diferencia   +/- Límite
--------------------------------------
 1 VS 2       -7.1016      7.9752
 1 VS 3       5.1863       7.7708
 2 VS 3      *12.2879     *8.6032
--------------------------------------

* Diferencia estadísticamente significativa.
```

Homocedasticidad de la opción Anova Un Factor.

```
Anova Un Factor. Homocedasticidad
=================================================================

Variable Respuesta:           FC2FC1
Variable Explicativa:         STATUS
Número de Casos:              40

Prueba C de Cochran: 0.4155     P-valor = 0.6692
Prueba de Bartlett:  1.3465     P-valor = 0.5101
```



Residuos vs Predicciones de la opción Anova Un Factor.

Anova Un Factor (a|y). Datos Agrupados

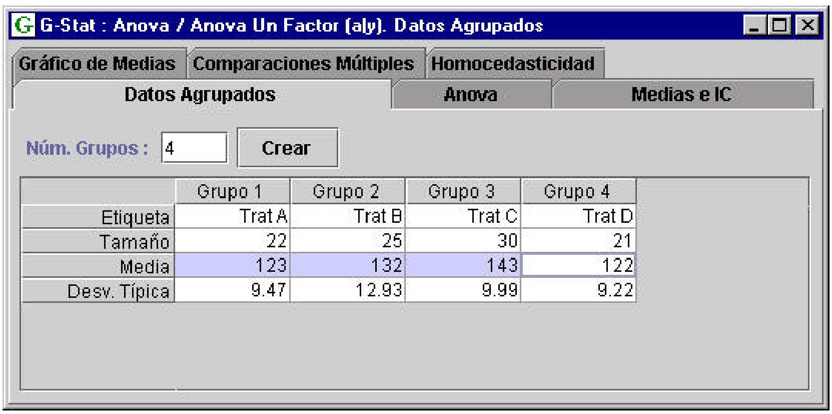
A partir de los datos resumidos de tamaño muestral, media y desviación típica de cada uno de los grupos, se puede realizar, sin el fichero de datos, la opción Anova Un Factor. Los fundamentos teóricos y la formulación son idénticos a los presentados en las opciones anteriores respectivas con datos a partir de fichero.

Manejo del programa

Datos Agrupados: Los datos necesarios son:

- Número de grupos.
- Etiqueta de cada grupo.
- Tamaño de cada grupo.
- Media de los grupos.
- Desviaciones típicas de cada grupo.

En la pantalla de entrada de datos agrupados, aparecen, por defecto, tres grupos con datos ficticios que hay que cambiar. Cada vez que se redefinen las dimensiones de la tabla aparecen los valores por defecto. Los datos, excepto la etiqueta, deben ser numéricos. No dejar filas o columnas con todos los valores faltantes o con todos los valores iguales a cero.



El resto de las pestañas se maneja igual que en la opción Anova anterior (sin datos agrupados).

Kruskal-Wallis (a|y)

Realiza la prueba no paramétrica de Kruskal-Wallis. Asimismo, facilita comparaciones múltiples a posteriori según el método de Dunn.

La prueba de Kruskal-Wallis es la más adecuada para comparar poblaciones cuyas distribuciones no son normales. Es la prueba no paramétrica análoga a la prueba paramétrica Anova. Incluso cuando las poblaciones son normales, este contraste funciona muy bien. También es adecuado cuando las desviaciones típicas de los diferentes grupos no son iguales entre sí, sin embargo, el Anova de un factor es muy robusto y sólo se ve afectado cuando las desviaciones típicas difieren en gran magnitud.

Las hipótesis de la prueba de Kruskal-Wallis son

H_0: Las k medianas son todas iguales

H_1: Al menos una de las medianas es diferente

La prueba de Kruskal-Wallis proporciona información en cuanto a la posible igualdad de medianas entre grupos y permite rechazar esta hipótesis si p-valor<0.05. En el caso de que se rechace la hipótesis nula de igualdad de medianas se puede determinar mediante comparaciones múltiples a posteriori, de qué grupo o grupos provienen esas diferencias.

Como paso previo al cálculo del estadístico de Kruskal-Wallis, a cada observación se le asigna el rango según el orden que ocupa la observación en el conjunto total de los datos, asignando el rango medio en caso de empates. A partir de estos rangos se define R_m como la suma de rangos de cada grupo m, m = 1,…,r, siendo r el número de grupos, y se calculan el valor medio de los rangos $E[R_m]$ y el rango medio $\overline{R}_m$ como

$$E[R_m] = \frac{n_m(n+1)}{2}$$

$$\overline{R}_m = \frac{R_m}{n_m}$$

Por último, el estadístico de contraste de Kruskal-Wallis H’ se calcula como:

$$H' = \frac{\frac{12}{n(n+1)} \sum_{m=1}^{r} \frac{1}{n_m} \left[R_m - E[R_m]\right]^2}{1 - \frac{\sum_{j=1}^{k} \left(d_j^3 - d_j\right)}{n^3 - n}}$$

siendo d_j el número de empates en j = 1,...,k con k el número de valores distintos de la variable respuesta, que sigue una distribución Chi-Cuadrado con r - 1 grados de libertad. Este programa proporciona un p-valor asintótico para esta opción.

Para realizar comparaciones múltiples a posteriori de los grupos se utiliza la prueba de Dunn. Para llevar a cabo esta prueba se establece la diferencia mínima entre los rangos medios de dos grupos para decretar la significación estadística a un nivel alfa determinado. Esta diferencia viene dada por

$$\left|\overline{R}_i - \overline{R}_j\right| > z_{1-\frac{\alpha}{k(k-1)}} \sqrt{\frac{1}{12(n-1)}\left[n\left(n^2-1\right) - \sum_{m=1}^{t}\left(d_m^3 - d_m\right)\right]\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}$$

donde $\overline{R}_i$ es el rango medio del grupo i, z es el valor de la distribución normal unilateral izquierda, alfa el nivel de significación, k el número de grupos, n el tamaño de la muestra, n_i el tamaño del grupo i, t el número de valores distintos de la variable respuesta y d_m el número de empates para el valor m de la variable respuesta.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa u ordinal y la *Variable explicativa* cualitativa que forma los grupos. La variable respuesta no puede ser constante. La variable explicativa debe tener dos o más grupos.

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Dispersión: Se muestran los datos de los casos para la variable respuesta según las categorías de la variable explicativa o factor. Así, para cada uno de los niveles del factor, que aparecen en el eje X, pueden verse los valores de la variable respuesta observados. Este gráfico permite tener una aproximación visual de cuál es el efecto del factor sobre la variable respuesta respecto de su media y de su dispersión.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Cajas: Gráfico de Cajas de la variable respuesta para los distintos valores de la variable explicativa. Ver menú Gráficos.

Kruskal-Wallis: Resultado de la prueba de Kruskal-Wallis.

Comparaciones Múltiples: Resultados de las comparaciones múltiples a posteriori por el método de Dunn para determinar de qué grupo o grupos provienen las diferencias detectadas en la prueba de Kruskal-Wallis. Este programa se basa en los resultados de las comparaciones dos a dos obtenidas. Mediante un asterisco se señalan los grupos que son diferentes y mediante un aspa se agrupan los grupos homogéneos o semejantes.

Opciones:

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

Se desea realizar la prueba de Kruskal-Wallis de la variable FC2FC1 con la variable Status como explicativa.

Resultados de la pestaña Kruskal-Wallis.

```
Kruskal-Wallis
======================================================================

Variable Respuesta:            FC2FC1
Variable Explicativa:          STATUS
Número de Casos:               40

----------------------------------------------------
 Grupos    N     Suma de Rangos Rm    Rango Medio
----------------------------------------------------
 1         17    338.0000             19.8824
 2         11    306.5000             27.8636
 3         12    175.5000             14.6250
----------------------------------------------------

Estadístico de Kruskal-Wallis (sin corrección por empates):  7.4424
Estadístico de Kruskal-Wallis (con corrección por empates):  7.4606
Grados de Libertad:  2
p-valor:             0.0240
```

Resultados de la prueba de comparaciones múltiples a posteriori de Dunn.

```
Kruskal-Wallis, Comparaciones Múltiples
======================================================================
Variable Respuesta:          EDAD
Variable Explicativa:        STATUS
Número de Casos:              40

Método: Dunn al 95.0%

----------------------------------------
                            Grupos
 EDAD   N    Rango Medio    Homogéneos
----------------------------------------
 2      11   15.3182         X
 1      17   20.5000         X
 3      12   25.2500         X
----------------------------------------


--------------------------------------
 Contraste   Diferencia   +/- Límite
--------------------------------------
 2 VS 1       5.1818       10.8234
 3 VS 1       -4.7500      10.5461
 3 VS 2       -9.9318      11.6757
--------------------------------------
* Diferencia estadísticamente significativa.
```

Anova Un Factor con Bloque (a bloque|y)

Realiza la prueba Anova un factor con bloque. Asimismo, facilita, entre otras técnicas asociadas a ella, las comparaciones múltiples a posteriori.

El Análisis de la Varianza de un factor con bloque compara medias entre distintos grupos y se basa en descomponer la variabilidad total en tres componentes: uno que se atribuye al hecho de pertenecer a un bloque u otro SCB, un segundo al de pertenecer a un grupo u otro SCG y un tercero con origen desconocido residual SCR. Se supone que la variable explicativa es un factor fijo y que la variable bloque es un factor aleatorio.

El programa proporciona la significación del factor y del bloque. El estadístico de contraste que se usa es el F para los grupos. La hipótesis nula de igualdad de medias se rechaza en el caso en el que p-valor<0.05, en caso contrario no hay evidencia suficiente para poder rechazarla. En el caso de que se rechace la hipótesis nula de igualdad de medias se puede determinar mediante comparaciones múltiples a posteriori, de qué grupo o grupos provienen esas diferencias

Si se denota por y al vector con los valores de la variable respuesta, con X_g a la matriz de n filas por ($1+r_g-1$) columnas con la primera columna todo de unos y las restantes (r_g-1) columnas, las asociadas a las variables dummy de la variable explicativa con r_g niveles, con X a la matriz de n filas por ($1+r_g-1+r_b-1$) columnas resultante de añadir (r_b-1) columnas a la matriz anterior X_g, asociadas a las variables dummy de la variable que forma los bloques con r_b niveles.

Se tiene que las expresiones para SCT (suma de cuadrados total), SCB (suma de cuadrados de la variable bloque), SCG (suma de cuadrados de la variable que forma los grupos), SCR (suma de cuadrados residual), GLT (grados de libertad total), GLB (grados de libertad de la variable bloque), GLG (grados de libertad de la variable que forma los grupos), GLR (grados de libertad residual), CMB (cuadrado medio de la variable bloque), CMG (cuadrado medio de la variable que forma los grupos), CMR (cuadrado medio residual), F_b (estadístico de contraste para la variable bloque) y F_g (estadístico de contraste para la variable que forma los grupos) son:

$$SCT = \left(y - \overline{y}\right)^t \left(y - \overline{y}\right)$$

$$SCG = SCT - \left(y - X_g b_g\right)^t \left(y - X_g b_g\right), \; b_g = \left(X_g^t X_g\right)^{-1} X_g^t y$$

$$SCR = \left(y - Xb\right)^t \left(y - Xb\right), \; b = \left(X^t X\right)^{-1} X^t y$$

$$SCB = SCT - SCG - SCR$$

$$GLT = n - 1, \; GLB = r_b - 1, \; GLG = r_g - 1, \; GLR = GLT - GLB - GLG$$

$$CMB = \frac{SCB}{GLB}, \; CMG = \frac{SCG}{GLG}, \; CMR = \frac{SCR}{GLR}$$

$$F_b = \frac{CMB}{CMR}, \; F_g = \frac{CMG}{CMR}$$

El estadístico de contraste F_b sigue una distribución F de Snedecor con grados de libertad del numerador GLB y grados de libertad del denominador GLR. El estadístico F_g sigue una distribución F de Snedecor con grados de libertad del numerador GLG y grados de libertad del denominador GLR.

Medias e IC

Se muestran para cada uno de los grupos, las medias de la variable cuantitativa, junto con su error estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña “Gráfico de Medias”.

En la pantalla de opciones se selecciona el método que se desee para el cálculo de los intervalos de las medias. Existen las siguientes posibilidades: ninguno, errores estándar, intervalos de confianza, intervalos LSD, intervalos HSD, intervalos Scheffé, intervalos de Bonferroni. Las fórmulas son análogas a las dadas en Anova / Anova / Medias e IC, aunque el cálculo de los errores estándar viene dado por

$$\sqrt{l \cdot \left(X^t X^* X\right)^{-1} \cdot l^t}$$

donde l es un vector para cada posible media con $(1+(r_g-1)+(r_b-1))$ coordenadas, donde la primera componente es uno, las componentes relativas a cada variable explicativa es 1 en la correspondiente variable dummy y cero en caso contrario, y las componentes relativas a la otra variable explicativa es el inverso del número de categorías de dicha variable explicativa.

Por ejemplo, si la variable que forma los grupos tiene dos categorías y la variable que forma los bloques también tiene dos categorías, para la primera categoría de la variable que forma los grupos el vector l es (1, 1, 1/2) y para la segunda categoría de la variable que forma los grupos el vector l es (1, 0, 1/2). Sólo se calculan medias para el factor fijo.

La matriz X^* es una matriz de dimensiones n por n formada por r_g bloques de dimensión n/r_g por n/r_g iguales a cero y bloques ABLOQ de dimensión n/r_g por n/r_g en la diagonal principal, siendo

$$ABLOQ = \begin{pmatrix} a+b & a & \dots & a \\ a & a+b & \dots & a \\ \dots & \dots & \dots & \dots \\ a & a & a & a+b \end{pmatrix}$$

con

$$a = \frac{CMB - CMR}{n / r_b} \quad y \quad b = CMR$$

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa, la *Variable explicativa* cualitativa y la *Variable bloque* también cualitativa. Para la aplicación de esta técnica es necesario que los datos estén balanceados para los dos factores, esto significa que todas las casillas de la posible interacción a*b tengan el mismo número de casos, si no es así el programa devuelve un mensaje de error.

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Dispersión: Se muestran los datos de los casos para la variable respuesta según las categorías de la variable explicativa o factor. Así, para cada uno de los niveles del factor, que aparecen en el eje X, pueden verse los valores de la variable respuesta observados. Este gráfico permite tener una aproximación visual de cuál es el efecto del factor sobre la variable respuesta respecto de su media y de su dispersión.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Cajas: Gráfico de Cajas de la variable respuesta para los distintos valores de la variable explicativa. Ver menú Gráficos.

Anova: Resultados del Análisis de la Varianza de un factor con bloque para la comparación de medias de los distintos grupos.

Medias e IC: Se muestran para cada uno de los grupos, las medias de la variable respuesta, junto con su error estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña "Gráfico de Medias".

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

Gráfico de Medias: Se muestran para cada uno de los grupos, las medias de la variable respuesta, junto con sus intervalos de confianza calculados según distintos métodos.

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Comparaciones Múltiples: Resultados de las pruebas a posteriori para determinar de qué grupo o grupos provienen las diferencias detectadas en el Anova. Este programa se basa en los resultados de las comparaciones dos a dos obtenidas. Mediante un asterisco se señalan los grupos que son diferentes y mediante un aspa se agrupan los grupos homogéneos o semejantes.

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

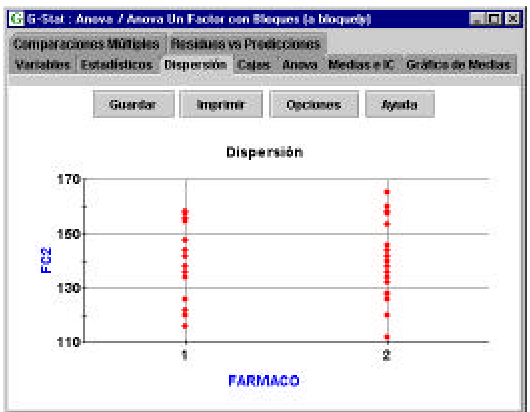
Residuos vs Predicciones: Se representan los residuales frente a los valores que predice el modelo. Este gráfico sirve para poder detectar falta de homocedasticidad (heterocedasticidad). La banda de residuos debería ser similar en dispersión y simetría a lo largo de todos los predichos.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Realizar la prueba Anova un factor con bloques con la variable FC2 como variable respuesta, Farmaco como variable explicativa y Sexo como variable bloque. Para que el diseño esté balanceado cambiar la variable Sexo=2, en los casos IB=27 y IB=30.

Variables en la opción Anova Un Factor con Bloques.



Dispersión en la opción Anova Un Factor con Bloques.

Resultados del Anova Un Factor con Bloques.

```
Anova Un Factor con Bloques
=============================================

Variable Respuesta:          FC2
Variable Explicativa:        FARMACO
Variable Bloque:             SEXO
Número de Casos:             40

Anova
--------------------------------------------------------------------
                   Suma de             Cuadrado
                   Cuadrados    G.L.   Medio         F-valor     p-valor
--------------------------------------------------------------------
 Entre Bloques     2190.4000    1      2190.4000     17.3711     0.0002
 Entre Grupos      22.5000      1      22.5000       0.1784      0.6752
 Residual          4665.5000    37     126.0946
--------------------------------------------------------------------
 Total (corr.)     6878.4000    39
```

Medias e IC de la opción Anova Un Factor con Bloques con los intervalos LSD.

```
Anova Un Factor con Bloques
=============================================

Variable Respuesta:          FC2
Variable Explicativa:        FARMACO
Variable Bloque:             SEXO
Número de Casos:             40

Tabla de medias con I.C. LSD al 95.0%
-----------------------------------------------------------
                                      Límite      Límite
 FARMACO   n    Media      E.E.       Inferior    Superior
-----------------------------------------------------------
 1         20   136.9500   7.6100     121.5306    152.3694
 2         20   138.4500   7.6100     123.0306    153.8694
-----------------------------------------------------------
 Total     2    137.7000
```

Comparaciones múltiples LSD de la opción Anova Un Factor con Bloques.

```
Anova Un Factor con Bloques. Comparaciones Múltiples
======================================================================

Variable Respuesta:          FC2
Variable Explicativa:        FARMACO
```

```
Variable Bloque:             SEXO
Número de Casos:             40

con I.C. LSD al 95.0%
----------------------------------------
                            Grupos
 FARMACO   n    Media       Homogéneos
----------------------------------------
 1         20   136.9500    X
 2         20   138.4500    X
----------------------------------------

 Contraste   Diferencia   +/- Límite
--------------------------------------
 1 VS 2       -1.5000      7.1949
--------------------------------------

* Diferencia estadísticamente significativa.
```

Friedman (a bloque|y)

Realiza la prueba no paramétrica de Friedman.

La prueba de Friedman es la análoga no paramétrica del Anova de un factor con muestras (bloques) pareadas, con lo que compara varias medianas en lugar de varias medias. Como toda técnica no paramétrica funciona con rangos en lugar de con los valores originales de la variable respuesta. Se basa en descomponer la variabilidad total en tres componentes: uno que se atribuye al hecho de pertenecer a un bloque u otro, un segundo al de pertenecer a un grupo u otro y un tercero con origen desconocido.

Se supone que la variable explicativa tiene k niveles y que hay b bloques. Dentro de cada bloque se asignan rangos a la variable respuesta, en el caso de empates se consideran los promedios de los órdenes de las observaciones empatadas. Para cada nivel j, j=1,...,k, se suman dichos rangos, obteniéndose R_j. Estos rangos verifican que

$$\sum_{j=1}^{k} R_j = \frac{k(k+1)}{2} b$$

El estadístico de contraste se calcula como

$$\chi_F^2 = b(k-1)W$$

que sigue una distribución Chi-Cuadrado con k-1 grados de libertad, siendo

$$W = \frac{12\left(\sum_{j=1}^{k} R_j^2\right) - 3b^2k(k+1)^2}{b^2k(k^2-1) - b\sum_{empates}\left(empates^3 - empates\right)}$$

Este programa proporciona un p-valor asintótico para esta opción. La hipótesis nula de igualdad de medianas en los grupos se rechaza en el caso en el que p-valor<0.05, en caso contrario no hay evidencia suficiente para poder rechazarla.

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa, la *Variable explicativa* cualitativa y la *Variable bloque* también cualitativa. La variable bloque está asociada al sujeto, que presenta datos longitudinales para cada una de las categorías del factor analizado. El factor puede ser, en muchas ocasiones, diferentes tiempos en un estudio de diseño longitudinal. Para la aplicación de esta técnica es requisito que los datos estén balanceados y que solamente haya una única observación para cada posible combinación de la variable explicativa y del bloque. Si no es así, el programa devuelve el siguiente mensaje de error:

```
ERROR :
La prueba de Friedman requiere datos balanceados y que solamente
haya una única observación para cada posible combinación de la
variable explicativa y del bloque.
```

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Dispersión: Se muestran los datos de los casos para la variable respuesta según las categorías de la variable explicativa o factor. Así, para cada uno de los niveles del factor, que aparecen en el eje X, pueden verse los valores de la variable respuesta observados. Este gráfico permite tener una aproximación visual de cuál es el efecto del factor sobre la variable respuesta respecto de su media y de su dispersión.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Cajas: Gráfico de Cajas de la variable respuesta para los distintos valores de la variable explicativa. Ver menú Gráficos.

Friedman: Resultados de la prueba de Friedman.

Se analizan tres tratamientos en seis sujetos. Se asume que el orden de administración no influye en la respuesta. Se pretende establecer si hay diferencia entre tratamientos. Los datos son los siguientes:

suj	trat	resp
1	1	9
1	2	6
1	3	5
2	1	11
2	2	8
2	3	7
3	1	9
3	2	7
3	3	6
4	1	12
4	2	9
4	3	9
5	1	7
5	2	4
5	3	4
6	1	19
6	2	16
6	3	17

Resultados de la estadística descriptiva.

```
Anova Friedman. Estadísticos
======================================================================

Variable Respuesta:          resp
Variable Explicativa:        trat
Variable Bloque:             sujeto
Número de Casos:             18
```

```
                                          Desviación
 Grupos    |   N     Media    Mediana     Típica    Mínimo  Máximo
-----------------------------------------------------------------
 1         |   6     11.1667  10.0000     4.2151    7.0000  19.0000
 2         |   6     8.3333   7.5000      4.1312    4.0000  16.0000
 3         |   6     8.0000   6.5000      4.7329    4.0000  17.0000
-----------------------------------------------------------------
 Total     |   18    9.1667   8.5000      4.3555    4.0000  19.0000
```

Resultados de la prueba de Friedman.

```
Friedman
======================================================================

Variable Respuesta:          resp
Variable Explicativa:        trat
Variable Bloque:             sujeto
Número de Casos:             18

 Grupos    N   Rango Medio
-----------------------------
 1         6   3.0000
 2         6   1.6667
 3         6   1.3333
-----------------------------

Coeficiente de Concordancia: 0.8485
Estadístico de Contraste:    10.1818
p-valor:                     0.0062
```

La prueba tiene en cuenta la información del bloque (sujeto). Este análisis realizado -erróneamente- por Kruskal-Wallis no hubiera detectado diferencias significativas entre tratamientos.

Anacova (ax|y)

Realiza la prueba del análisis de la covarianza, Anacova, con y sin interacciones. Asimismo, facilita, entre otras técnicas asociadas a ella, las comparaciones múltiples a posteriori.

El Anacova o Análisis de la Varianza de un factor con covariable se basa en descomponer la variabilidad total SCT en tres componentes: uno que se atribuye a la covariable SCC, un segundo al hecho de pertenecer a un grupo u otro SCE y un tercero con origen desconocido residual SCR.

Si se denota por y al vector con los valores de la variable respuesta, con X_c a la matriz de n filas por dos columnas con la primera columna todo de unos y la segunda columna con los valores de la covariable y con X a la matriz de n filas por (1+1+r-1) columnas resultante de añadir (r-1) columnas a la matriz anterior X_c, asociadas a las variables dummy de la variable explicativa con r niveles.

Se tiene que las expresiones para SCT (suma de cuadrados total), SCC (suma de cuadrados de la covariable), SCE (suma de cuadrados entre los grupos), SCR (suma de cuadrados residual), GLT (grados de libertad total), GLC (grados de libertad de la covariable), GLE (grados de libertad entre los grupos), GLR (grados de libertad residual), CMC (cuadrado medio de la covariable), CME (cuadrado medio entre grupos), CMR (cuadrado medio residual), F_c (estadístico de contraste para la covariable) y F_e (estadístico de contraste para la variable que forma los grupos) son:

$$SCT = \left(y - \overline{y}\right)^t \left(y - \overline{y}\right)$$

$$SCC = SCT - \left(y - X_c b_c\right)^t \left(y - X_c b_c\right), \; b_c = \left(X_c^t X_c\right)^{-1} X_c^t y$$

$$SCR = \left(y - Xb\right)^t \left(y - Xb\right), \; b = \left(X^t X\right)^{-1} X^t y$$

$$SCE = SCT - SCC - SCR$$

$$GLT = n - 1, \; GLC = 1, \; GLE = r - 1, \; GLR = GLT - GLC - GLE$$

$$CMC = \frac{SCC}{GLC}, \; CME = \frac{SCE}{GLE}, \; CMR = \frac{SCR}{GLR}$$

$$F_c = \frac{CMC}{CMR}, \; F_e = \frac{CME}{CMR}$$

El estadístico de contraste F_c sigue una distribución F de Snedecor con grados de libertad del numerador GLC y grados de libertad del denominador GLR. El estadístico F_e sigue una distribución F de Snedecor con grados de libertad del numerador GLE y grados de libertad del denominador GLR. La hipótesis nula de igualdad de medias se rechaza en el caso en el que F_e tenga un p-valor<0.05, en caso contrario no hay evidencia suficiente para poder rechazarla. En el caso de que se rechace la hipótesis nula de igualdad de medias se puede determinar mediante comparaciones múltiples a posteriori, de qué grupo o grupos provienen esas diferencias.

Medias e IC

Se muestran para cada uno de los grupos, las medias de la variable cuantitativa, junto con su error estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña "Gráfico de Medias".

En la pantalla de opciones se selecciona el método que se desee para el cálculo de los intervalos de las medias. Existen las siguientes posibilidades: ninguno, errores estándar, intervalos de confianza, intervalos LSD, intervalos HSD, intervalos Scheffé, intervalos Bonferroni. Las fórmulas son análogas a las dadas en Anova / Anova / Medias e IC, aunque el cálculo de los errores estándar viene dado por

$$\sqrt{CMR \cdot l \cdot \left(X^t X\right)^{-1} \cdot l^t}\ ,$$

donde l es un vector para cada posible media con (1+(r-1)+1) coordenadas, donde la primera componente es uno, las siguientes (r-1) componentes son 1 si es la correspondiente variable dummy y cero en caso contrario, y la última componente es la media de la covariable. Por ejemplo, si la variable explicativa tiene tres categorías, para la primera categoría el vector l es (1, 1, 0, media de la covariable), para la segunda categoría el vector l es (1, 0, 1, media de la covariable) y para la tercera categoría es (1, 0, 0, media de la covariable).

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa, la *Variable explicativa* cualitativa (factor) y la *Covariable* cuantitativa. Ni la variable respuesta ni la covariable pueden ser constantes. La variable explicativa debe tener dos o más grupos.

Estadísticos: Estadísticos de la variable respuesta según las categorías de la variable explicativa. Se detallan en Cuantitativa (y).

Dispersión: Se muestran los datos de los casos para la variable respuesta según las categorías de la variable explicativa o factor. Así, para cada uno de los niveles del factor, que aparecen en el eje X, pueden verse los valores de la variable respuesta observados. Este gráfico permite tener una aproximación visual de cuál es el efecto del factor sobre la variable respuesta respecto de su media y de su dispersión.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Cajas: Gráfico de Cajas de la variable respuesta para los distintos valores de la variable explicativa. Ver menú Gráficos.

Anova: Resultado del Análisis de la Varianza de un factor con covariable para la comparación de medias de los distintos grupos en presencia de una covariable.

Medias e IC: Se muestran para cada uno de los grupos, las medias de la variable respuesta, junto con su error estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña "Gráfico de Medias".

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

Gráfico de Medias: Se muestran para cada uno de los grupos, las medias de la variable respuesta, junto con sus intervalos de confianza calculados según distintos métodos.

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Comparaciones Múltiples: Resultados de las pruebas a posteriori para determinar de qué grupo o grupos provienen las diferencias detectadas en el Anova. Este programa se basa en los resultados de las comparaciones dos a dos obtenidas. Mediante un asterisco se señalan los grupos que son diferentes y mediante un aspa se agrupan los grupos homogéneos o semejantes.

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

Residuos vs Grupos: Se representan los residuales para cada uno de los grupos. Los residuales se calculan como la diferencia entre el valor observado y el valor que predice el modelo. Este gráfico sirve para poder detectar falta de homocedasticidad (heterocedasticidad). Los grupos deberían tener dispersiones de los residuos similares.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Residuos vs Predicciones: Se representan los residuales frente a los valores que predice el modelo. Este gráfico sirve para poder detectar falta de homocedasticidad (heterocedasticidad). La banda de residuos debería ser similar en dispersión y simetría a lo largo de todos los predichos.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Residuos vs Registros: Se representan los residuales frente al número de registro en la base de datos. Este gráfico no debería mostrar ningún patrón no aleatorio. Así, para cada una de las unidades de estudio del fichero de trabajo (las filas o registros), que aparecen en el eje X, se muestra el valor del residuo del modelo estimado. Si la nube de puntos no muestra ninguna pauta o patrón (el rango de oscilación de los residuos no depende del valor de la fila, no se observa periodos continuados de residuos crecientes o decrecientes, etc.), se tendrá una indicación gráfica de que los residuos son incorrelados: el error cometido para una unidad de estudio no depende de los errores cometidos para las unidades inmediatamente anteriores.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Aplicar la prueba Anacova a la variable respuesta FC2, con la variable Farmaco como variable explicativa y FC1 como covariable.

Resultados de la estadística descriptiva.

Anacova. Estadísticos

Variable Respuesta: FC2
Variable(s) Explicativa(s): FARMACO, FC1
Número de Casos: 40

Grupo	N	Media	Mediana	Desviación Típica	Mínimo	Máximo
1	20	136.9500	137.0000	13.3435	116.0000	158.0000
2	20	138.4500	137.0000	13.5199	112.0000	165.0000
Total	40	137.7000	137.0000	13.2804	112.0000	165.0000
FC1	40	75.9500	74.0000	9.5379	62.0000	96.0000

Anova de la opción Anacova.

Anacova

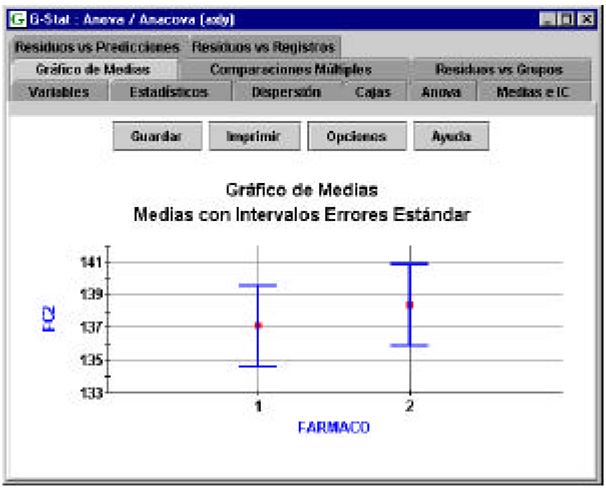
Variable Respuesta: FC2
Variable(s) Explicativa(s): FARMACO, FC1
Número de Casos: 40

Anova

	Suma de Cuadrados	G.L.	Cuadrado Medio	F-valor	p-valor
Covariable	2310.9613	1	2310.9613	18.7858	0.0001
Entre Grupos	15.8266	1	15.8266	0.1287	0.7219
Residual	4551.6121	37	123.0165		
Total (corr.)	6878.4000	39			

Medias e IC de la opción Anacova con los intervalos LSD.

```
Anacova. Medias e I.C.
======================================================================

Variable Respuesta:           FC2
Variable(s) Explicativa(s):   FARMACO, FC1
Número de Casos:              40

Tabla de Medias con I.C. LSD al 95.0%
--------------------------------------------------------------
                                          Límite      Límite
 FARMACO    N     Media       E. E.       Inferior    Superior
--------------------------------------------------------------
 1          20    137.0709    2.4802      133.5174    140.6244
 2          20    138.3291    2.4802      134.7756    141.8826
--------------------------------------------------------------
 Total      40    137.7000
```

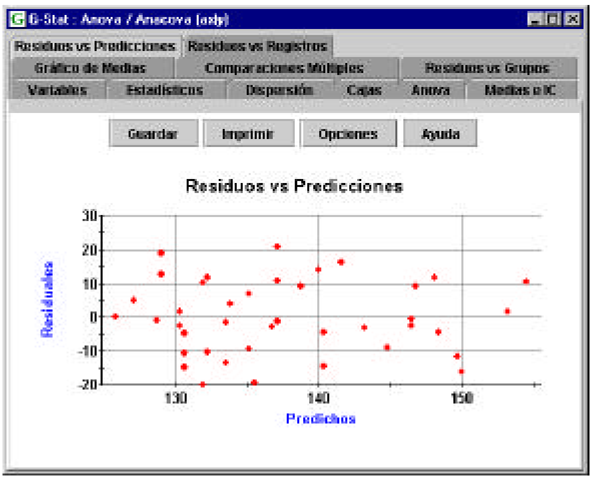


Gráficos de Medias con Intervalos errores estándar

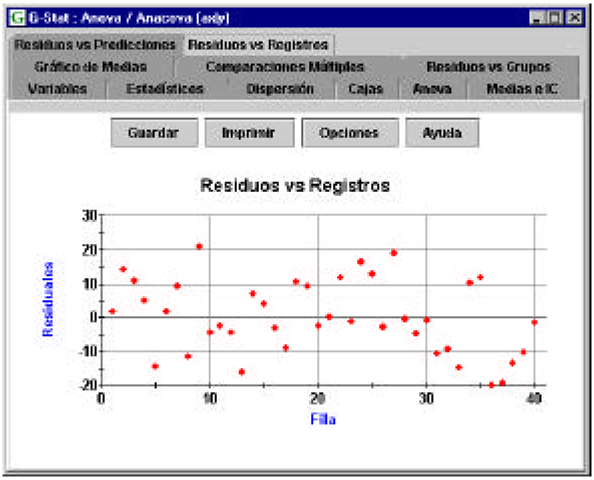
Comparaciones Múltiples por Bonferroni de la opción Anacova.

```
Anacova. Comparaciones Múltiples
======================================================================

Variable Respuesta:           FC2
Variable(s) Explicativa(s):   FARMACO, FC1
Número de Casos:              40
```

```
con I.C. Bonferroni al 95.0%
----------------------------------------
                            Grupos
 FARMACO   n    Media       Homogéneos
----------------------------------------
 1         20   137.0709    X
 2         20   138.3291    X
----------------------------------------

 Contraste   Diferencia   +/- Límite
--------------------------------------
 1 VS 2       -1.2582      7.1075
--------------------------------------

* Diferencia estadísticamente significativa.
```



Residuos vs Predicciones de la opción Anacova.



Residuos vs Registros de la opción Anacova.

No se observan tendencias, patrones, ni variaciones en la dispersión de los residuos.

Anova Dos Factores (ab|y)

Realiza la prueba de Anova con dos factores con y sin interacción. Asimismo, facilita, entre otras técnicas asociadas a ella, las comparaciones múltiples a posteriori.

El Análisis de la Varianza de dos factores se basa en descomponer la variabilidad total SCT en tres componentes: uno para un factor SCA, un segundo para el otro factor SCB y un tercero con origen desconocido SCR.

Se denota por y al vector con los valores de la variable respuesta, con X_a a la matriz de n filas por $(1+r_a-1)$ columnas con la primera columna todo de unos y las restantes (r_a-1) columnas, las asociadas a las variables dummy de la primera variable explicativa con r_a niveles, con X a la matriz de n filas por $(1+r_a-1+r_b-1)$ columnas resultantes de añadir (r_b-1) columnas a la matriz anterior X_a, asociadas a las variables dummy de la segunda variable explicativa con r_b niveles.

Se tiene que las expresiones para SCT (suma de cuadrados total), SCA (suma de cuadrados de la primera variable explicativa), SCB (suma de cuadrados de la segunda variable explicativa, SCR (suma de cuadrados residual), GLT (grados de libertad total), GLA (grados de libertad de la primera variable explicativa), GLB (grados de libertad de la segunda variable explicativa), GLR (grados de libertad residual), CMA (cuadrado medio de la primera variable explicativa), CMB (cuadrado medio de la segunda variable explicativa), CMR (cuadrado medio residual), F_a (estadístico de contraste para la primera variable explicativa) y F_b (estadístico de contraste para la segunda variable explicativa) son:

$$SCT = \left(y - \overline{y}\right)^t \left(y - \overline{y}\right)$$

$$SCA = SCT - \left(y - X_a b_a\right)^t \left(y - X_a b_a\right), \ b_a = \left(X_a^t X_a\right)^{-1} X_a^t y$$

$$SCR = \left(y - Xb\right)^t \left(y - Xb\right), \ b = \left(X^t X\right)^{-1} X^t y$$

$$SCB = SCT - SCA - SCR$$

$$GLT = n - 1, \ GLA = r_a - 1, \ GLB = r_b - 1, GLR = GLT - GLA - GLB$$

$$CMA = \frac{SCA}{GLA}, \ CMB = \frac{SCB}{GLB}, \ CMR = \frac{SCR}{GLR}$$

$$F_a = \frac{CMA}{CMR}, \ F_b = \frac{CMB}{CMR}$$

El estadístico de contraste F_a sigue una distribución F de Snedecor con grados de libertad del numerador GLA y grados de libertad del denominador GLR. El estadístico F_b sigue una distribución F de Snedecor con grados de libertad del numerador GLB y grados de libertad del denominador GLR. Para cada factor, la hipótesis nula de igualdad de medias se rechaza en el caso en el que su F tenga un $p\text{-valor} < 0.05$, en caso contrario no hay evidencia suficiente para poder rechazarla. En el caso de que se rechace la hipótesis nula de igualdad de medias se puede determinar mediante comparaciones múltiples a posteriori, de qué grupo o grupos provienen esas diferencias.

Se incluye también la posibilidad de contemplar la interacción entre los dos factores, que a efectos computacionales es como un nuevo factor que se crea a partir de la combinación de los niveles de los dos factores. La hipótesis nula del factor de interacción está relacionada con la ausencia de interacción. Si su F tiene un $p\text{-valor} < 0.05$ se rechaza la ausencia de interacción.

Medias e IC

Se muestran para cada uno de los grupos, las medias de la variable cuantitativa, junto con su error estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña “Gráfico de Medias”.

En la pantalla de opciones se selecciona el método que se desee para el cálculo de los intervalos de las medias. Existen las siguientes posibilidades: ninguno, errores estándar, intervalos de confianza, intervalos LSD, intervalos HSD, intervalos Scheffé, intervalos Bonferroni. Las fórmulas son análogas a las dadas en Anova / Anova / Medias e IC, aunque el cálculo de los errores estándar viene dado por

$$\sqrt{CMR \cdot l \cdot \left(X^t X\right)^{-1} \cdot l^t}\ ,$$

donde l es un vector para cada posible media con $(1+(r_a-1)+(r_b-1))$ coordenadas, donde la primera componente es uno, las componentes relativas a cada variable explicativa es 1 en la correspondiente variable dummy y cero en caso contrario, y las componentes relativas a la otra variable explicativa es el inverso del número de categorías de dicha variable explicativa.

Por ejemplo, si la primera variable explicativa tiene tres categorías y la segunda variable explicativa tiene dos categorías, para la primera categoría de la primera variable explicativa el vector l es (1, 1, 0, 1/2), para la segunda categoría de la primera variable explicativa el vector l es (1, 0, 1, 1/2), para la tercera categoría de la primera variable explicativa es (1, 0, 0, 1/2), para la primera categoría de la segunda variable explicativa el vector l es (1, 1/3, 1/3, 1) y para la segunda categoría de la segunda variable explicativa el vector l es (1, 1/3, 1/3, 0).

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y las dos *Variables explicativas* cualitativas que forman los grupos. La variable respuesta no puede ser constante. Las variables explicativas deben tener dos o más grupos.

Estadísticos: Estadísticos de la variable respuesta para cada categoría de las variables explicativas y sus combinaciones. Se detallan en Cuantitativa (y).

Dispersión: Se muestran los datos de los casos para la variable respuesta según las categorías de cada variable explicativa o factor. Así, para cada uno de los niveles del factor, que aparecen en el eje X, pueden verse los valores de la variable respuesta observados. Este gráfico permite tener una aproximación visual de cuál es el efecto del factor sobre la variable respuesta respecto de su media y de su dispersión.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Anova: Resultados del Análisis de la Varianza de dos factores para la comparación de medias de los distintos grupos.

Opciones:

- Incluir o no el término interacción en el modelo.

Medias e IC: En esta tabla se muestran para cada uno de los grupos de los dos factores, las medias de la variable respuesta, junto con su error estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña "Gráfico de Medias".

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

- Incluir o no el término interacción en el modelo.

Gráfico de Medias: Se muestran para cada uno de los grupos de los dos factores, las medias de la variable respuesta, junto con sus intervalos de confianza calculados según distintos métodos.

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

- Incluir o no el término interacción en el modelo.

Comparaciones Múltiples: Resultados de las pruebas a posteriori para determinar de qué grupo o grupos de los dos factores provienen las diferencias detectadas en el Anova. Este programa se basa en los resultados de las comparaciones dos a dos obtenidas. Mediante un asterisco se señalan los grupos que son diferentes y mediante un aspa se agrupan los grupos homogéneos o semejantes.

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza entre >0 y <100.

- Incluir o no el término interacción en el modelo.

Gráfico Interacciones: El objetivo de este gráfico es detectar la posible interacción entre los factores, que se presenta en el caso de que en el gráfico no haya paralelismo entre las rectas. En caso de que haya interacción, la interpretación de la influencia de los factores no es directa. La propia combinación de los efectos de cada uno de los factores que forman parte del estudio, puede ser el resultado de la variabilidad de la variable respuesta, lo que se conoce como interacción. Gráficamente la interacción de factores se refleja mediante la ausencia de paralelismo de las rectas que unen las medias.

Aplicar la prueba del Anova dos factores para analizar la variable FC2FC1 como variable respuesta empleando las variables Farmaco y Status como variables explicativas.

Resultados descriptivos.

Anova Dos Factores. Estadísticos

Variable Respuesta: FC2FC1
Variable(s) Explicativa(s): STATUS, FARMACO
Número de Casos: 40

STATUS	N	Media	Mediana	Desviación Típica	Mínimo	Máximo
3	12	56.1667	53.0000	10.5644	42.0000	74.0000
2	11	68.4545	69.0000	7.8913	60.0000	82.0000
1	17	61.3529	59.0000	11.1183	42.0000	82.0000
Total	40	61.7500	61.0000	10.9772	42.0000	82.0000

FARMACO	N	Media	Mediana	Desviación Típica	Mínimo	Máximo
2	20	62.3500	62.0000	9.6206	44.0000	78.0000
1	20	61.1500	59.5000	12.4108	42.0000	82.0000
Total	40	61.7500	61.0000	10.9772	42.0000	82.0000

STATUS FARMACO	N	Media	Mediana	Desviación Típica	Mínimo	Máximo
2,2	6	66.8333	65.5000	7.3869	60.0000	78.0000
2,1	5	70.4000	70.0000	8.8769	60.0000	82.0000
1,2	8	62.2500	62.0000	9.7651	48.0000	76.0000
1,1	9	60.5556	59.0000	12.7388	42.0000	82.0000
3,2	6	58.0000	58.0000	10.8074	44.0000	74.0000
3,1	6	54.3333	52.0000	10.9848	42.0000	74.0000
Total	40	61.7500	61.0000	10.9772	42.0000	82.0000

Resultados del Anova con interacción.

Anova Dos Factores

Variable Respuesta: FC2FC1
Variable(s) Explicativa(s): STATUS, FARMACO
Número de Casos: 40

Anova

	Suma de Cuadrados	G.L.	Cuadrado Medio	F-valor	p-valor
STATUS	871.2237	2	435.6119	3.9590	0.0285
FARMACO	7.1657	1	7.1657	0.0651	0.8001
STATUS*FARMACO	80.0217	2	40.0108	0.3636	0.6978
Residual	3741.0889	34	110.0320		
Total (corr.)	4699.5000	39			

Medias e IC con los errores estándar de los subgrupos sin interacción.

```
Anova Dos Factores, Medias e I.C.
======================================================================

Variable Respuesta:          FC2FC1
Variable(s) Explicativa(s):  STATUS, FARMACO
Número de Casos:             40

Tabla de Medias con Intervalos Errores Estándar
----------------------------------------------------------
                                    Límite      Límite
           n    Media      E.E.     Inferior    Superior
----------------------------------------------------------
 Total     40   61.9868

 STATUS
 1         17   61.3779    2.5006   58.8773     63.8785
 2         11   68.4160    3.1099   65.3061     71.5259
 3         12   56.1667    2.9741   53.1926     59.1407

 FARMACO
 1         20   61.5628    2.3387   59.2241     63.9015
 2         20   62.4109    2.3142   60.0967     64.7251
```

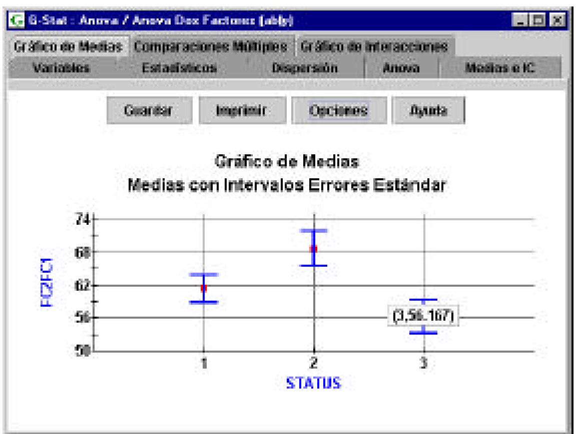


Gráfico de medias con Intervalos errores estándar para el factor Status.

Comparaciones múltiples con la prueba de Scheffé en un modelo con interacción.

```
Anova Dos Factores. Comparaciones Múltiples
======================================================================

Variable Respuesta:           FC2FC1
Variable(s) Explicativa(s):   STATUS, FARMACO
Número de Casos:              40

Modelo con Interacción

con I.C. Scheffé al 95.0%
--------------------------------------
                          Grupos
 STATUS   n    Media      Homogéneos
--------------------------------------
 3        12   56.1667    X
 1        17   61.4028    XX
 2        11   68.6167     X
--------------------------------------

 Contraste   Diferencia   +/- Límite
--------------------------------------
 1 VS 2       -7.2139      10.4229
 1 VS 3       5.2361       10.1306
 2 VS 3      *12.4500     *11.2320
--------------------------------------
* Diferencia estadísticamente significativa.
```

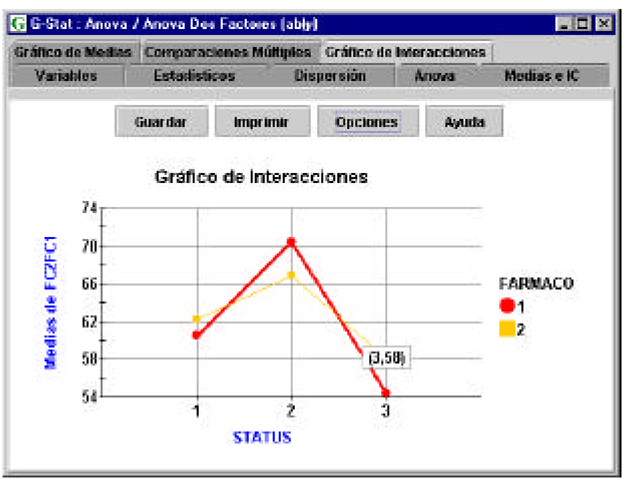


Gráfico de medias con indicación de las interacciones.

Anova Factorial (abc|y)

Realiza la prueba Anova factorial para tres factores con y sin interacciones dobles y triples. Asimismo, facilita, entre otras técnicas asociadas a ella, las comparaciones múltiples a posteriori.

El Análisis de la Varianza Factorial (tres factores) se basa en descomponer la variabilidad total SCT en cuatro componentes: para el primer factor SCA, para el segundo factor SCB, para el tercer factor SCC y un cuarto con origen desconocido SCR.

Se denota por y al vector con los valores de la variable respuesta, con X_a a la matriz de n filas por $(1+r_a-1)$ columnas con la primera columna todo de unos y las restantes (r_a-1) columnas, las asociadas a las variables dummy de la primera variable explicativa con r_a niveles, con X_{ab} a la matriz de n filas por $(1+r_a-1+r_b-1)$ columnas resultante de añadir (r_b-1) columnas a la matriz anterior X_a, asociadas a las variables dummy de la segunda variable explicativa con r_b niveles, con X a la matriz de n filas por $(1+r_a-1+r_b-1+r_c-1)$ columnas resultante de añadir (r_c-1) columnas a la matriz anterior X_{ab}, asociadas a las variables dummy de la tercera variable explicativa con r_c niveles.

Se tiene que las expresiones para SCT (suma de cuadrados total), SCA (suma de cuadrados de la primera variable explicativa), SCB (suma de cuadrados de la segunda variable explicativa), SCC (suma de cuadrados de la tercera variable explicativa), SCR (suma de cuadrados residual), GLT (grados de libertad total), GLA (grados de libertad de la primera variable explicativa), GLB (grados de libertad de la segunda variable explicativa), GLC (grados de libertad de la tercera variable explicativa), GLR (grados de libertad residual), CMA (cuadrado medio de la primera variable explicativa), CMB (cuadrado medio de la segunda variable explicativa), CMC (cuadrado medio de la tercera variable explicativa), CMR (cuadrado medio residual), F_a (estadístico de contraste para la primera variable explicativa), F_b (estadístico de contraste para la segunda variable explicativa) y F_c (estadístico de contraste para la tercera variable explicativa) son:

$$SCT = \left(y - \overline{y}\right)^t\left(y - \overline{y}\right)$$

$$SCA = SCT - \left(y - X_a b_a\right)^t\left(y - X_a b_a\right), \quad b_a = \left(X_a^t X_a\right)^{-1} X_a^t y$$

$$SCB = SCT - SCA - \left(y - X_{ab} b_{ab}\right)^t\left(y - X_{ab} b_{ab}\right), \quad b_{ab} = \left(X_{ab}^t X_{ab}\right)^{-1} X_{ab}^t y$$

$$SCR = \left(y - Xb\right)^t\left(y - Xb\right), \quad b = \left(X^t X\right)^{-1} X^t y$$

$$SCC = SCT - SCA - SCB - SCR$$

$$GLT = n - 1\,,\ \ GLA = r_a - 1\,,\ \ GLB = r_b - 1\,,\ \ GLC = r_c - 1$$
$$GLR = GLT - GLA - GLB - GLC$$

$$CMA = \frac{SCA}{GLA}\,,\quad CMB = \frac{SCB}{GLB}\,,\quad CMC = \frac{SCC}{GLC}\,,\quad CMR = \frac{SCR}{GLR}$$

$$F_a = \frac{CMA}{CMR}\,,\ \ F_b = \frac{CMB}{CMR}\,,\ \ F_c = \frac{CMC}{CMR}$$

El estadístico de contraste F_a sigue una distribución F de Snedecor con grados de libertad del numerador GLA y grados de libertad del denominador GLR. El estadístico F_b sigue una distribución F de Snedecor con grados de libertad del numerador GLB y grados de libertad del denominador GLR. El estadístico F_c sigue una distribución F de Snedecor con grados de libertad del numerador GLC y grados de libertad del denominador GLR. Para cada factor, la hipótesis nula de igualdad de medias se rechaza en el caso en el que su F tenga un p-valor<0.05, en caso contrario no hay evidencia suficiente para poder rechazarla. En el caso de que se rechace la hipótesis nula de igualdad de medias se puede determinar mediante comparaciones múltiples a posteriori, de qué grupo o grupos provienen esas diferencias.

Se incluye también la posibilidad de contemplar las interacciones entre dos factores o incluso entre los tres factores, que a efectos computacionales son como un nuevo factor que se crea a partir de la combinación de los niveles de los dos o tres factores.

Medias e IC

En la tabla correspondiente se muestran para cada uno de los grupos, las medias de la variable cuantitativa, junto con su error estándar y sus intervalos confidenciales calculados según distintos métodos. Estos intervalos confidenciales se pueden representar gráficamente con la pestaña "Gráfico de Medias".

En la pantalla de opciones se selecciona el método que se desee para el cálculo de los intervalos de las medias. Existen las siguientes posibilidades: ninguno, errores estándar, intervalos de confianza, intervalos LSD, intervalos HSD, intervalos Scheffé, intervalos Bonferroni. Las fórmulas son análogas a las dadas en Anova / Anova / Medias e IC, aunque el cálculo de los errores estándar viene dado por

$$\sqrt{CMR \cdot I \cdot \left(X^t X\right)^{-1} \cdot I^t}$$

donde l es un vector para cada posible media con $(1+(r_a-1)+(r_b-1)+(r_c-1))$ coordenadas, donde la primera componente es uno, las componentes relativas a cada variable explicativa es 1 en la correspondiente variable dummy y cero en caso contrario, y las componentes relativas a las otras variables explicativas son el inverso del número de categorías de la correspondiente variable explicativa.

Por ejemplo, si la primera variable explicativa tiene tres categorías, la segunda variable explicativa tiene dos categorías y la tercera variable explicativa tiene dos categorías, para la primera categoría de la primera variable explicativa el vector l es (1, 1, 0, 1/2, 1/2), para la segunda categoría de la primera variable explicativa el vector l es (1, 0, 1, 1/2, 1/2), para la tercera categoría de la primera variable explicativa es (1, 0, 0, 1/2, 1/2), para la primera categoría de la segunda variable explicativa el vector l es (1, 1/3, 1/3, 1, 1/2), para la segunda categoría de la segunda variable explicativa el vector l es (1, 1/3, 1/3, 0, 1/2), para la primera categoría de la tercera variable explicativa el vector l es (1, 1/3, 1/3, 1/2, 1) y para la segunda categoría de la tercera variable explicativa el vector l es (1, 1/3, 1/3, 1/2, 0).

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y las tres *Variables explicativas* cualitativas que forman los grupos. La variable respuesta no puede ser constante. Las variables explicativas deben tener dos o más grupos.

Estadísticos: Estadísticos de la variable respuesta para cada categoría de las variables explicativas y sus combinaciones. Se detallan en Cuantitativa (y).

Dispersión: Se muestran los datos de los casos para la variable respuesta según las categorías de cada variable explicativa o factor. Así, para cada uno de los niveles del factor, que aparecen en el eje X, pueden verse los valores de la variable respuesta observados. Este gráfico permite tener una aproximación visual de cuál es el efecto del factor sobre la variable respuesta respecto de su media y de su dispersión.

Opciones:

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

Anova: Resultados del Análisis de la Varianza de dos factores para la comparación de medias de los distintos grupos.

Opciones:

- Incluir o no los términos de interacción doble o triple en el modelo. Si se señala el análisis de la interacción triple, también se incluirán las dobles.

Opciones de Anova en Anova Factorial.

Medias e IC: En esta tabla se muestran para cada uno de los grupos de los tres factores , las medias de la variable respuesta, junto con su error estándar y sus intervalos de confianza calculados según distintos métodos. Estos intervalos de confianza se pueden representar gráficamente con la pestaña "Gráfico de Medias".

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

- Incluir o no los términos de interacción doble o triple en el modelo. Si se señala el análisis de la interacción triple, también se incluirán las dobles.

Gráfico de Medias: Se muestran para cada uno de los grupos de los tres factores, las medias de la variable respuesta, junto con sus intervalos de confianza calculados según distintos métodos.

Opciones:

- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

- La cabecera, el título del eje X y del eje Y, el mínimo, máximo e incremento del eje Y.

- Incluir o no los términos de interacción doble o triple en el modelo. Si se señala el análisis de la interacción triple, también se incluirán las dobles.

Comparaciones Múltiples: Resultados de las pruebas a posteriori para determinar de qué grupo o grupos de los tres factores provienen las diferencias detectadas en el Anova. Este programa se basa en los resultados de las comparaciones dos a dos obtenidas. Mediante un asterisco se señalan los grupos que son diferentes y mediante un aspa se agrupan los grupos homogéneos o semejantes.

Opciones:

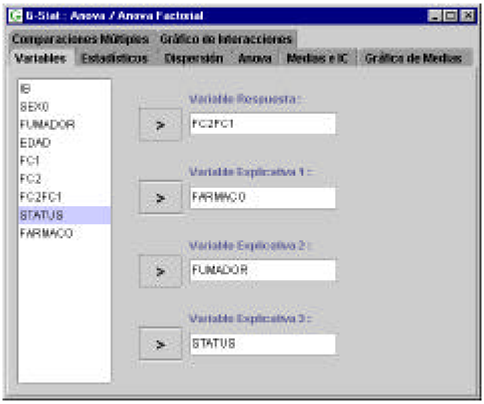
- Método: Ver opción Anova un Factor.

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

- Incluir o no los términos de interacción doble o triple en el modelo. Si se señala el análisis de la interacción triple, también se incluirán las dobles.

Gráfico Interacciones: El objetivo de este gráfico es detectar la posible interacción entre los factores, que se presenta en el caso de que en el gráfico no haya paralelismo entre las rectas. En caso de que haya interacción, la interpretación de la influencia de los factores no es directa. La propia combinación de los efectos de cada uno de los factores que forman parte del estudio, puede ser el resultado de la variabilidad de la variable respuesta, lo que se conoce como interacción. Gráficamente la interacción de factores se refleja mediante la ausencia de paralelismo de las rectas que unen las medias.

Realizar la prueba Anova Factorial de la variable FC2FC1 con Farmaco, Fumador y Status como variables explicativas.



Variables en la opción Anova Factorial.

Estadísticos de la opción Anova Factorial (sólo se muestran una interacción doble y la triple).

```
Anova Factorial. Estadísticos
======================================================================

Variable Respuesta:          FC2FC1
Variable(s) Explicativa(s):  FARMACO, FUMADOR, STATUS
Número de Casos:              40
. . . . . . . . .

 FUMADOR                                Desviación
 STATUS     N  Media      Mediana     Típica        Mínimo     Máximo
----------------------------------------------------------------------
 2,3        9  54.4444    52.0000     10.3816       42.0000    74.0000
 2,2        6  67.3333    65.0000     8.4538        60.0000    78.0000
 2,1        9  60.7778    58.0000     10.6040       46.0000    82.0000
 1,3        3  61.3333    58.0000     11.3725       52.0000    74.0000
 1,2        5  69.8000    69.0000     7.8867        62.0000    82.0000
 1,1        8  62.0000    67.0000     12.3751       42.0000    76.0000
----------------------------------------------------------------------
 Total      40 61.7500    61.0000     10.9772       42.0000    82.0000
----------------------------------------------------------------------
```

FARMACO FUMADOR STATUS	N	Media	Mediana	Desviación Típica	Mínimo	Máximo
2,1,2	3	67.6667	69.0000	5.1316	62.0000	72.0000
2,1,1	4	65.0000	68.0000	12.0554	48.0000	76.0000
1,1,3	3	61.3333	58.0000	11.3725	52.0000	74.0000
1,1,2	2	73.0000	73.0000	12.7279	64.0000	82.0000
1,1,1	4	59.0000	61.0000	13.7113	42.0000	72.0000
2,2,3	6	58.0000	58.0000	10.8074	44.0000	74.0000
2,2,2	3	66.0000	60.0000	10.3923	60.0000	78.0000
1,2,3	3	47.3333	48.0000	5.0332	42.0000	52.0000
2,2,1	4	59.5000	58.0000	7.5498	52.0000	70.0000
1,2,2	3	68.6667	70.0000	8.0829	60.0000	76.0000
1,2,1	5	61.8000	59.0000	13.3866	46.0000	82.0000
Total	40	61.7500	61.0000	10.9772	42.0000	82.0000

Tabla del Anova Factorial con interacciones dobles.

Anova Factorial

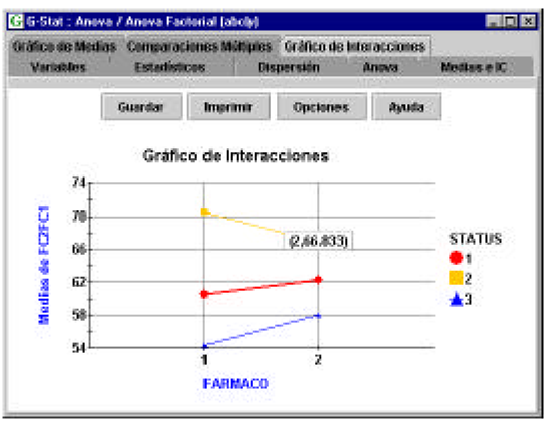
Variable Respuesta: FC2FC1
Variable(s) Explicativa(s): FARMACO, FUMADOR, STATUS
Número de Casos: 40

Anova

	Suma de Cuadrados	G.L.	Cuadrado Medio	F-valor	p-valor
FARMACO	14.4000	1	14.4000	0.1274	0.7237
FUMADOR	187.4568	1	187.4568	1.6581	0.2077
STATUS	764.0682	2	382.0341	3.3793	0.0475
INTERACCIONES					
A*B	5.4190	1	5.4190	0.0479	0.8282
A*C	120.2221	2	60.1110	0.5317	0.5930
B*C	216.3783	2	108.1891	0.9570	0.3955
Residual	3391.5556	30	113.0519		
Total (corr.)	4699.5000	39			

Comparaciones múltiples por Bonferroni de la opción Anova factorial con interacciones dobles para la variable Status.

```
Anova Factorial. Comparaciones Múltiples
======================================================================

Variable Respuesta:           FC2FC1
Variable(s) Explicativa(s):   FARMACO, FUMADOR, STATUS
Número de Casos:               40

Modelo con interacciones 2 niveles

con I.C. Bonferroni al 95.0%
--------------------------------------
                          Grupos
 STATUS   n    Media      Homogéneos
--------------------------------------
 3        12   60.6786    X
 1        17   61.3810    X
 2        11   68.6468    X
--------------------------------------

 Contraste   Diferencia   +/- Límite
--------------------------------------
 1 VS 2       -7.2659      10.5089
 1 VS 3       0.7024       12.7744
 2 VS 3       7.9683       13.8128
--------------------------------------

* Diferencia estadísticamente significativa.
```



Interacciones de la opción Anova factorial para Farmaco por Status.

Menú Multivariante

Activar la opción **Multivariante** del menú principal o mediante Alt+M. Este menú contiene las opciones necesarias para realizar distintas regresiones multivariantes. En Regresión Múltiple se tiene que varias variables explicativas cuantitativas (xz) intentan explicar una variable respuesta cuantitativa (y). En Regresión Logística se tiene que varias variables explicativas cuantitativas (xz) intentan explicar una variable respuesta dicotómica o binaria (b). En Regresión de Cox se tiene que varias variables explicativas cuantitativas (xz) intentan explicar una variable respuesta censurada (y cens). En las tres regresiones si se desea utilizar variables explicativas cualitativas de k categorías, es necesario generar previamente k-1 variables ficticias y asignarles los códigos binarios correspondientes. Todas las variables explicativas han de ser numéricas, inclusive las de naturaleza dicotómica.

La forma de generación de k-1 variables ficticias a partir de una variable cualitativa con k categorías es la siguiente:

Variable Cualitativa	Ficticia 1	Ficticia 2	...	Ficticia k-1
Categoría 1	0	0	...	0
Categoría 2	1	0	...	0
Categoría 3	0	1	...	0
...	...	...	...	...
Categoría k	0	0	...	1

Este paso no es automático en G-Stat. La forma de proceder es la siguiente: insertar k-1 variables con sus nombres correspondientes, copiar (mediante CTRL+C y CTRL+V) k-1 veces la variable cualitativa que se quiere transformar en “dummy” en las variables insertadas y recodificar dichas variables según el esquema de la tabla anterior.

Para el estudio de las interacciones entre variables explicativas hay que crear previamente las variables de interacción como producto aritmético de dichas variables mediante el menú de Utilidades / Transformación. Posteriormente se tratan como una variable explicativa más.

Regresión Múltiple (xz|y)

Realiza la regresión lineal múltiple que modeliza una variable respuesta cuantitativa a partir de varias variables explicativas cuantitativas.

El modelo que se asume para describir la relación entre el conjunto de r variables explicativas y la variable respuesta y es

$$y = X\beta + \varepsilon$$

donde y es un vector de dimensiones n por 1, X es una matriz de dimensiones n por (1+r) con la primera columna igual a uno, β es el vector de parámetros del modelo de dimensiones (r+1) por 1 y ε es el vector de residuos de dimensiones n por 1.

El vector de parámetros β se estima por el vector de coeficientes b a través del método de mínimos cuadrados

$$b = \left(X^t X\right)^{-1} X^t y$$

A partir del modelo se calculan los valores predichos mediante

$$\hat{y} = Xb$$

por lo que los residuos estimados son

$$e = \hat{\varepsilon} = y - \hat{y} = y - Xb$$

Mediante el vector de residuos estimados se calcula la desviación típica estimada de los residuos s con

$$s = \sqrt{\frac{\hat{\varepsilon}^t \hat{\varepsilon}}{n - (r + 1)}}$$

El vector de errores estándar de los coeficientes b se estima a través de la raíz cuadrada de los elementos de la diagonal principal de la matriz Cov(b) de varianzas-covarianzas de b dada por

$$Cov(b) = s^2 \left(X^t X\right)^{-1}$$

La significación de cada variable se lee en cada uno de los p-valores asociados a cada coeficiente, y se calcula a través del estadístico t que resulta de dividir el coeficiente entre su error estándar.

La tabla del Anova muestra mediante la F del Modelo si el modelo ajusta a los datos. En dicha tabla intervienen SCT (suma de cuadrados total), SCM (suma de cuadrados del modelo), SCR (suma de cuadrados residual), GLT (grados de libertad total), GLM (grados de libertad del modelo), GLR (grados de libertad de los residuos), CMM (cuadrado medio del modelo), CMR (cuadrado medio residual), F del modelo y p del modelo, que se calculan de la forma siguiente:

$$SCT = y^t y - n\bar{y}^2$$

$$SCM = \hat{y}^t \hat{y} - n\bar{\hat{y}}^2$$

$$SCR = \hat{\varepsilon}^t \hat{\varepsilon} = SCT - SCM$$

$$GLT = n - 1, \; GLM = r, \; GLR = n - (r + 1) = GLT - GLM$$

$$CMM = \frac{SCM}{GLM}, \; CMR = \frac{SCR}{GLR}$$

$$F = \frac{CMM}{CMR}$$

que sigue una distribución F de Snedecor con grados de libertad del numerador GLM y grados de libertad del denominador GLR.

El coeficiente R^2 de determinación suministra el porcentaje de información de la variable respuesta explicado por el modelo mediante

$$R^2 = \frac{SCM}{SCT}$$

Un ajuste de R^2 teniendo en cuenta el número de variables, ya que a mayor número de variables se corresponde mayor R^2, es

$$R^2 \text{ ajustado} = 1 - \left(1 - R^2\right)\frac{GLT}{GLR}$$

El coeficiente de variación se calcula mediante la expresión

$$100 \cdot \frac{s}{\bar{y}}$$

La detección de correlación en los residuales la da el estadístico de Durbin-Watson mediante la expresión

$$\frac{\sum_{j=2}^{n}\left(\hat{\varepsilon}_j - \hat{\varepsilon}_{j-1}\right)^2}{\sum_{j=1}^{n}\hat{\varepsilon}_j^2}$$

Coeficientes e IC

Para cada coeficiente b_j del vector b se calcula su intervalo confidencial como

$$b_j \pm t_{1-\alpha/2;\ n-(r+1)} \cdot EE\left[b_j\right]$$

El coeficiente Factor Incremento de la Varianza FIV (“Variance Inflation Factor” = VIF) permite detectar la presencia de multicolinealidad y se calcula como

$$VIF_j = \frac{1}{1 - R_j^2}$$

siendo R_j^2 el coeficiente de determinación múltiple en una regresión con variable respuesta la variable x_j y variables explicativas el resto de variables x. Por tanto, la multicolinealidad se da cuando las correlaciones entre variables independientes son elevadas, lo cual no es conveniente porque la estimación del modelo puede no ser posible.

Predicciones

Se muestran los residuos, los residuos estudentizados y las predicciones. Los residuos vienen dados por e, con

$$e = \hat{\varepsilon} = y - \hat{y} = y - Xb$$

Los residuos estudentizados calculan los residuos “jacknife” $r_{(-i)}$ mediante

$$r_{(-i)} = \frac{e_i}{\sqrt{s_{(-i)}^2 \cdot \left(1 - h_i\right)}}$$

donde $s_{(-i)}$ es la desviación típica estimada de los residuos cuando se suprime la observación del individuo i y h_i es el elemento i de la diagonal de la matriz H de dimensiones n por n dada por

$$H = X\left(X^t X\right)^{-1} X^t$$

La matriz H recibe el nombre de “hat matrix” ya que

$$\hat{y} = Hy$$

Los residuos "jacknife" pueden requerir elevados recursos computacionales para ficheros de más de 1000 casos.

Para las predicciones de valores individuales se utiliza

$$\text{Pred}_i \pm t_{1-\alpha/2;\, n-(r+1)} \cdot s\sqrt{\frac{1}{\text{npred}} + h_i}$$

siendo npred el número usado para las predicciones.

Para las predicciones de valores medios se utiliza

$$\text{Pred}_i \pm t_{1-\alpha/2;\, n-(r+1)} \cdot s\sqrt{h_i}$$

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa y la(s) V*ariable(s) explicativa(s)* cuantitativas. Las variables no deben tener varianza cero.

Estadísticos: Estadísticos de todas las variables seleccionadas. Se detallan en Cuantitativa (y). Sólo se tienen en cuenta los registros completos para todas las variables analizadas.

Correlaciones: Se muestran los coeficientes de correlación de Pearson y de Spearman para cada par de variables. La significación dada por el p-valor (entre paréntesis) está contrastando la hipótesis nula de que el coeficiente de correlación poblacional es cero.

Modelo: Se muestra la ecuación del modelo, la tabla del Anova, el coeficiente R^2 y el estadístico de Durbin-Watson, entre otros.

Opciones:

- *S*i se quiere incluir o no el término constante.

- Selección de variables: incluir todas, aplicar un procedimiento paso a paso hacia adelante o hacia detrás.

- Nivel de significación p para entrar y para salir en el proceso secuencial de selección del modelo: las variables van entrando en el modelo si realmente lo mejoran más allá de lo que podría deberse al azar (p-para-entrar) y pueden salir si no mejoran el modelo significativamente (con relación a p-para-salir).

- Nivel de confianza: Por defecto es 95%, pero también son habituales 90% y 99%. El nivel de confianza debe ser >0 y <100.

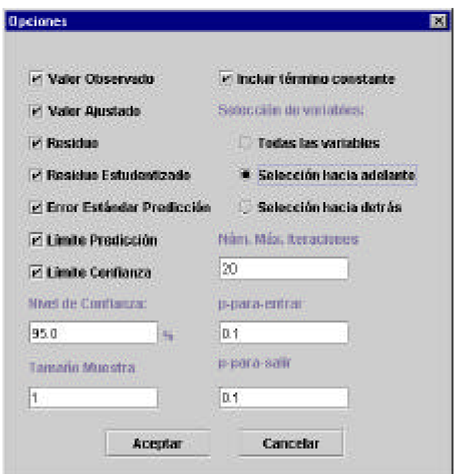
- Número máximo de iteraciones que se permite.

Por defecto, el programa tiene marcadas las opciones: incluido el término constante, incluidas todas las variables, p-para-entrar y p-para-salir con valor 0.1 y el número máximo de iteraciones es 20.

Aunque el algoritmo utilizado es robusto en la mayoría de situaciones, no se asegura la convergencia en todos los casos.

Coeficientes e IC: Se muestran para cada coeficiente de la regresión múltiple su error estándar, su intervalo de confianza y su coeficiente Factor Incremento de la Varianza (FIV). El nivel de confianza se puede modificar en la ventana de opciones. En las opciones el valor por defecto es 95%. El nivel de confianza debe ser >0 y <100. En opciones se debe redefinir el modelo tal y como se definió en la pestaña Modelo. Las ventanas de opciones no están conectadas entre ventanas, por lo que cualquier cambio en una pestaña debe realizarse nuevamente en las demás si se desean resultados coherentes.

Predicciones: Se presentan predicciones e intervalos de confianza partir de las variables explicativas. En opciones se debe redefinir el modelo tal y como se definió en la pestaña Modelo.



Se desea modelizar mediante regresión lineal múltiple la variable FC2FC1 en función de las variables Edad y FC1.

Resultados de la matriz de correlaciones de Pearson y su significación.

```
Regresión Múltiple. Correlaciones
======================================================================

Variable Respuesta:                FC2FC1
Variable(s) Explicativa(s):        EDAD, FC1
Número de Casos:                   40

r de Pearson
(Significación)
           FC2FC1      EDAD        FC1
---------------------------------------------
 FC2FC1    1.0000      -0.9393     -0.1676
                       (0.0001)    (0.3012)

 EDAD      -0.9393     1.0000      0.1848
           (0.0001)                (0.2537)

 FC1       -0.1676     0.1848      1.0000
           (0.3012)    (0.2537)
----------------------------------------------------------------------
```

Resultados del modelo, incluyendo todas las variables, con la significación de los coeficientes.

```
Regresión Lineal Múltiple. Modelo
======================================================================

Variable Respuesta:                FC2FC1
Variable(s) Explicativa(s):        EDAD, FC1
Número de Casos:                   40
---------------------------------------------------------
             Coef.      E.E.       t-valor     p-valor
---------------------------------------------------------
 CONSTANTE   222.1198   10.1887    21.8005     0.0001E-18
 EDAD        -7.1784    0.4382     -16.3821    0.0002E-14
 FC1         0.0070     0.0661     0.1064      0.9158
---------------------------------------------------------
```

Anova

Variabilidad	Suma de Cuadrados	G.L.	Cuadrado Medio	F-valor	p-valor
Modelo	4146.0952	2	2073.0476	138.6015	0.0007E-14
Residual	553.4048	37	14.9569		
Total (corr.)	4699.5000	39			
Total (corr.)	4699.5000	39			

r cuadrado (coficiente de determinación)	88.2242 %
r cuadrado (ajustado)	87.5876 %
Desviación Típica de los Residuos	3.8674
Coeficiente de variación	6.2630 %
Error Absoluto Medio	2.9014
Durbin-Watson	1.5536

Intervalos de confianza de los coeficientes y valores FIV.

Regresión Lineal Múltiple. Coeficientes e I.C.

Variable Respuesta: FC2FC1
Variable(s) Explicativa(s): EDAD, FC1
Número de Casos: 40

Coeficientes e I.C. al 95.00%

	Coef.	E.E.	Límite Inferior	Límite Superior	Factor Incremento Varianza (FIV)
CONSTANTE	222.1198	10.1887	201.4756	242.7640	
EDAD	-7.1784	0.4382	-8.0662	-6.2906	1.0353
FC1	0.0070	0.0661	-0.1268	0.1409	1.0353

Estimados y análisis de residuales de algunos casos.

Regresión Lineal Múltiple. Predicciones

Variable Respuesta: FC2FC1
Variable(s) Explicativa(s): EDAD, FC1
Número de Casos: 40

```
-----------------------------------------------------------------
 Valor        Valor                     Límite Conf.     Límite Conf.
 Observado    Predicho    Residuo       Inferior 95.00   Superior 95.00
-----------------------------------------------------------------
 59.0000      60.5629     -1.5629       57.6299          63.4960
 76.0000      79.1002     -3.1002       76.5655          81.6350
 72.0000      67.6007     4.3993        66.1653          69.0362
 70.0000      68.9380     1.0620        66.6549          71.2211
 46.0000      46.8115     -0.8115       44.6013          49.0217
 66.0000      63.9412     2.0588        62.1386          65.7439
 68.0000      61.2245     6.7755        59.2003          63.2488
. . . . .
```

Regresión Logística (xz|b)

Realiza la regresión logística que modeliza una variable respuesta dicotómica o binaria (relacionada con la ocurrencia de un suceso) a partir de varias variables explicativas cuantitativas. Conviene codificar la variable respuesta Y con unos y ceros, de forma que el código uno se asocie al suceso de interés. El modelo de regresión logística para p variables explicativas x_1, x_2, ..., x_p es el siguiente:

$$\mathrm{Prob}(y_i = 1) = \frac{1}{1 + \exp(-(b_0 + b_1 x_1 + \ldots + b_p x_p))},$$

donde b_j está asociado a la variable explicativa j-ésima y b_0 es un coeficiente que no está asignado a ninguna variable, también llamado constante o “intercepta”.

Estos coeficientes se determinan maximizando la función de verosimilitud de la muestra. El procedimiento de maximización se realiza por el método iterativo de Newton-Raphson, disminuyendo en cada iteración el “–2Log Likelihood”, hasta alcanzar su valor mínimo en el modelo final. Este método además de proporcionar los coeficientes del modelo, también facilita sus errores estándar.

A partir de los coeficientes b_i y de los errores estándar $EE(b_i)$ se construye el estadístico de contraste de Wald con

$$\mathrm{Wald}_j = \left(\frac{b_j}{EE(b_j)}\right)^2$$

que sigue una Chi-Cuadrado con 1 grado de libertad.

El coeficiente de correlación parcial R se calcula como

$$R = \sqrt{\frac{Wald_j - 2}{-2LL(0)}}$$

donde –2LL(0) es menos dos veces el logaritmo neperiano de la función de verosimilitud en el paso inicial (paso cero). El signo de R es el signo del coeficiente. Si Wald es menor que dos, se considera R igual a cero.

El contraste de ajuste del modelo (Diferencia de Likelihood) contrasta si el decremento en el "–2Log Likelihood" entre el modelo nulo (sin incluir ninguna variable explicativa) y el modelo final es significativo, siendo el número de grados de libertad del estadístico de contraste el número de variables explicativas). No rechazar la hipótesis nula equivale a aceptar que conjuntamente las variables explicativas no son relevantes en el modelo (sus coeficientes son todos nulos).

La medida de efecto de cada variable se calcula a través del Odds Ratio que viene dado por la exponencial del coeficiente. Los intervalos de confianza del 95% de cada Odds Ratio vienen dados por

$$\left(\exp\left(b_j - 1.96\,EE\left(b_j\right)\right),\ \exp\left(b_j + 1.96\,EE\left(b_j\right)\right)\right)$$

Un OR significativamente mayor que uno indica que un incremento en la variable explicativa se asocia a un incremento en el Odds de la variable respuesta. Igualmente, un OR significativamente menor que uno indica que un incremento en la variable explicativa se asocia a un decremento en el Odds de la variable respuesta. El OR se considerará significativamente diferente de 1 si es significativamente distinto de cero el coeficiente correspondiente en el modelo

Para realizar la tabla de clasificación, se calcula para cada individuo i la probabilidad estimada de ocurrencia que viene dada por

$$Prob_i = \frac{1}{1 + \exp\left(-\left(b_0 + b_1 x_{1i} + \ldots + b_p x_{pi}\right)\right)}$$

Si $Prob_i >=$ Punto de corte que define el usuario entonces se predice "el valor que se esté modelizando en la ventana de opciones del modelo", en caso contrario se predice "el valor que no se esté modelizando ". La tabla de clasificación resulta de cruzar la variable respuesta observada con la predicha. El porcentaje de casos correctamente pronosticados puede entenderse como un estimador de la calidad de ajuste del modelo, aunque éste siempre estará sesgado hacia valores altos debido a que los mismos casos que han permitido estimar el modelo son los que se están pronosticando.

Manejo del programa

Variables: Se identifica la *Variable respuesta* dicotómica o binaria (b) y la(s) V*ariable(s) explicativa(s)* cuantitativas. Las variables no deben tener varianza cero.

Estadísticos: Estadísticos de las variables explicativas para cada nivel de la variable respuesta. Se detallan en Cuantitativa (y). Sólo se tienen en cuenta los registros completos para todas las variables analizadas.

Modelo: Se muestra el proceso iterativo de disminución del estadístico "-2Log Likelihood", la ecuación del modelo, los "Odds ratio" y la tabla de clasificación.

Opciones:

- *S*i se quiere incluir o no el término constante.

- Código de ocurrencia que se quiere modelizar de la variable respuesta dicotómica.

- Selección de variables: incluir todas, aplicar un procedimiento paso a paso hacia adelante o hacia detrás.

- Valor del punto de corte que se utilizará para generar la tabla de clasificación que proporciona el modelo.

- Nivel de significación p para entrar y para salir en el proceso secuencial de selección del modelo: las variables van entrando en el modelo si realmente lo mejoran más allá de lo que podría deberse al azar (p-para-entrar) y pueden salir si no mejoran el modelo significativamente (con relación a p-para-salir).

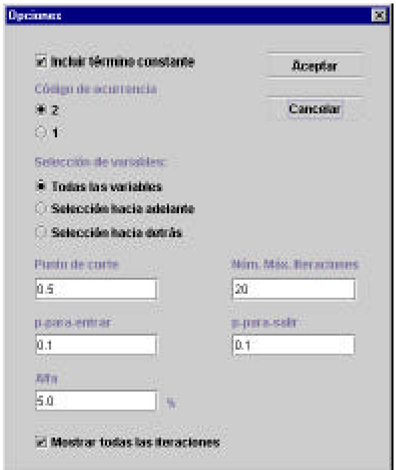
- Valor alfa para construir los intervalos de confianza para el OR asociado a cada variable explicativa del modelo. En las opciones el valor por defecto de alfa es 5% que corresponde a un IC del 95%. Alfa debe ser >0 y <100.

- Número máximo de iteraciones que se permite.

Por defecto, el programa tiene marcadas las opciones: incluido el término constante, código de ocurrencia igual al primer valor en el fichero de datos para la variable respuesta, incluidas todas las variables, punto de corte 0.5, p-para-entrar y p-para-salir con valor 0.1, alfa de un 5% y el número máximo de iteraciones es 20.

En presencia de separación o cuasiseparación los estimadores de máxima verosimilitud no existen. No obstante, se presentan los resultados que se deducen de la última iteración. En estos casos la validez del modelo es cuestionable.

Aunque el algoritmo utilizado es robusto en la mayoría de situaciones, no se asegura la convergencia en todos los casos.



Menú de opciones de la regresión logística en Modelo.

Predicciones: Para cada caso se presentan la predicción por el modelo y su residuo. En opciones se debe redefinir el modelo tal y como se definió en la pestaña Modelo. Las ventanas de opciones no están conectadas en todo el programa, por lo que cualquier cambio en una pestaña debe realizarse nuevamente en las demás si se desean resultados coherentes.

Se desea modelizar la variable Fumador, con código de ocurrencia igual a 2, mediante las variables Sexo, Edad y FC2FC1.

Estadísticos en la regresión logística.

```
Regresión Logística. Estadísticos
======================================================================

Variable Respuesta:             FUMADOR
Variable(s) Explicativa(s):     SEXO, EDAD, FC2FC1
Número de Casos:                40
```

```
Variable=SEXO
                                      Desviación
 FUMADOR   N     Media     Mediana    Típica      Mínimo   Máximo
-------------------------------------------------------------------
 2         24    1.5000    1.5000     0.5108      1.0000   2.0000
 1         16    1.3750    1.0000     0.5000      1.0000   2.0000
-------------------------------------------------------------------
 Total     40    1.4500    1.0000     0.5038      1.0000   2.0000

Variable=EDAD
                                       Desviación
 FUMADOR   N     Media      Mediana    Típica      Mínimo     Máximo
---------------------------------------------------------------------
 2         24    22.5833    22.7500    1.3871      19.5000    25.6000
 1         16    22.1625    21.9500    1.5209      20.0000    25.4000
---------------------------------------------------------------------
 Total     40    22.4150    22.5500    1.4380      19.5000    25.6000

Variable=FC2FC1
                                       Desviación
 FUMADOR   N     Media      Mediana    Típica      Mínimo     Máximo
---------------------------------------------------------------------
 2         24    60.0417    59.5000    10.8847     42.0000    82.0000
 1         16    64.3125    67.0000    10.9527     42.0000    82.0000
---------------------------------------------------------------------
 Total     40    61.7500    61.0000    10.9772     42.0000    82.0000
```

Resultados de la opción Modelo en la regresión logística.

```
Regresión Logística. Modelo
=====================================================================

 Variable Respuesta:                  FUMADOR
 Valor modelizado (ocurrencia):       FUMADOR = 2
 Variable(s) Explicativa(s):          SEXO, EDAD, FC2FC1
 Número de Casos:                     40
 Número de Casos con FUMADOR = 2 :    24
 Número de Casos con FUMADOR = 1 :    16

El modelo ha convergido satisfactoriamente

 Función Logaritmo de la Verosimilitud LL (Log Likelihood)
------------------------------------------------------------
 -2 Log Likelihood = 53.8409 (Modelo Nulo)
 -2 Log Likelihood = 49.8110
 -2 Log Likelihood = 49.7624
 -2 Log Likelihood = 49.7623
 -2 Log Likelihood = 49.7623 (Modelo Final)
```

```
Ajuste del Modelo (Diferencia de Likelihood)
-----------------------------------------------------------------
 Chi-Cuadrado =    4.0786;    G.L.    3;    p-valor =    0.2531

R Cuadrado del Modelo:
-------------------------
 Cox-Snell =     0.0969
 Nagelkerke =    0.1310

Regresión Logística:
---------------------------------------------------------------------
 Variable    Coef.     E.E.      Wald     G.L.    p-valor    R
---------------------------------------------------------------------
 SEXO        1.0968    0.7826    1.9640   1       0.1611     0.0000
 EDAD        -0.6149   0.7140    0.7415   1       0.3892     0.0000
 FC2FC1      -0.1331   0.0980    1.8448   1       0.1744     0.0000
 CONSTANTE   20.8624   21.5571   0.9366   1       0.3332     0.0000

 Variable   OR       IC95.0%inf   IC95.0%sup
----------------------------------------------
 SEXO       2.9946   0.6459       13.8843
 EDAD       0.5407   0.1334       2.1915
 FC2FC1     0.8754   0.7225       1.0607

Tabla de Clasificación para FUMADOR (Punto de corte = 0.50))
---------------------------------------------------------
               |   Predicción    |        |
 Observados    |   2             |   1    |   % Correcto
---------------------------------------------------------
 2             |   19            |   5    |   79.1667 %
 1             |   11            |   5    |   31.2500 %
---------------------------------------------------------
               |                 |        |   60.0000 %

Índices de Diagnóstico
---------------------------------------
 Valor Predictivo Positivo =    0.6333
 Valor Predictivo Negativo =    0.5000
 Sensibilidad              =    0.7917
 Especificidad             =    0.3125
 Indice de Youden          =    0.1042
```

Regresión de Cox (xz|y cens)

Realiza la regresión de Cox que modeliza una variable respuesta cuantitativa censurada por la derecha en función de una o varias variables explicativas cuantitativas. Adicionalmente se muestran las gráficas de las funciones de supervivencia por el método de Kaplan-Meier.

Kaplan-Meier

Gráfico de las curvas de supervivencia para cada categoría de cada posible variable explicativa cualitativa mediante el procedimiento de Kaplan-Meier. Estas curvas se interpretan como la probabilidad de “sobrevivir” a un tiempo dado y permiten identificar cuartiles de interés (p.ej. la mediana). Si las curvas se cortan estamos ante una situación de modelos no proporcionales.

La forma de calcular la estimación de Kaplan-Meier para cada grupo m, m=1, ..., r, es

$$S(t_{mj}) = \prod_{i=1}^{j} \frac{n_{mi} - d_{mi}}{n_{mi}}$$

siendo $t_{m1}<t_{m2}<...<t_{mk}$ los tiempos ordenados de ocurrencia de suceso para el grupo m, n_{mj} el número de individuos en riesgo del grupo m en t_{mj} y d_{mj} el número de individuos del grupo m que experimentan el suceso en t_{mj}.

Si las curvas se cortan estamos ante una situación de modelos no proporcionales y se desaconseja el uso de la regresión de Cox como modelo de estimación.

Regresión de Cox

La regresión de Cox se utiliza cuando se quiere analizar la variable respuesta “tiempo hasta que ocurre un suceso” en función de varias variables explicativas. La particularidad de esta técnica es que trabaja con datos censurados, es decir con información parcial.

El modelo de regresión de Cox para p variables explicativas x_1, x_2, ..., x_p es

$$h(t, x_1, x_2, \ldots, x_p) = h_0(t) \cdot \exp(b_1 x_1 + \ldots + b_p x_p)$$

donde $h(t, x_1, x_2, ..., x_p)$ es la función de riesgo para un individuo con perfil (x_1, x_2, ..., x_p) a tiempo t y $h_0(t)$ representa la función de riesgo basal para un individuo con $x_1=0$, $x_2=0$, ..., $x_p=0$ y b_j está asociado a la variable explicativa j-ésima.

Estos coeficientes se determinan maximizando la función de verosimilitud de la muestra. El procedimiento de maximización se realiza por el método iterativo de Newton-Raphson, disminuyendo en cada iteración el “–2Log Likelihood”, hasta alcanzar su valor mínimo en el modelo final. Este método además de proporcionar los coeficientes del modelo, también facilita sus errores estándar.

A partir de los coeficientes b_j y de los errores estándar $EE(b_j)$ se construye el estadístico de contraste de Wald con

$$\mathrm{Wald}_j = \left(\frac{b_j}{EE(b_j)}\right)^2$$

que sigue una Chi-Cuadrado con 1 grado de libertad.

El coeficiente de correlación parcial R se calcula como

$$R = \sqrt{\frac{\mathrm{Wald}_j - 2}{-2LL(0)}}$$

donde –2LL(0) es menos dos veces el logaritmo neperiano de la función de cuasi-verosimilitud en el paso inicial (paso cero). El signo de R es el signo del coeficiente. Si Wald es menor que dos, se considera R igual a cero.

El contraste de ajuste del modelo (Diferencia de Likelihood) contrasta si el decremento en el "–2Log Likelihood" entre el modelo nulo (sin incluir ninguna variable explicativa) y el modelo final es significativo, siendo el número de grados de libertad del estadístico de contraste el número de variables explicativas). No rechazar la hipótesis nula equivale a aceptar que conjuntamente las variables explicativas no son relevantes en el modelo (sus coeficientes son todos nulos).

La medida de efecto de cada variable se calcula a través del "Hazard Ratio" (HR) que viene dado por la exponencial del coeficiente. Los intervalos de confianza del 95% de cada "Hazard Ratio" vienen dados por

$$\left(\exp\left(b_j - 1.96EE\left(b_j\right)\right), \ \exp\left(b_j + 1.96EE\left(b_j\right)\right)\right)$$

Un HR significativamente mayor que uno indica que un incremento en la variable explicativa se asocia a un incremento en el riesgo y, por tanto, a una disminución en la supervivencia. Igualmente, un HR significativamente menor que uno indica que un incremento en la variable explicativa se asocia a un decremento en el riesgo y, por tanto, a un aumento en la supervivencia. El HR se considerará significativamente diferente de 1 si es significativamente distinto de cero el coeficiente correspondiente en el modelo

Manejo del programa

Variables: Se identifica la *Variable respuesta* cuantitativa, la *Variable de censura* dicotómica y la(s) *Variable(s) explicativa(s)* cuantitativas. Las variables no deben tener varianza cero.

Estadísticos: Estadísticos de las variables explicativas para cada nivel de la variable censura. Se detallan en Cuantitativa (y). Sólo se tienen en cuenta los registros completos para todas las variables analizadas.

Kaplan-Meier Tabla: Para cada tiempo exacto se incluye, según el método de Kaplan-Meier, la probabilidad de supervivencia y la mediana del tiempo de supervivencia.

Opciones:

- Se identifica código para datos censurados.

- Obtener las probabilidades de supervivencia para todos los casos o estratificarlas por grupos de una variable explicativa del modelo.

Kaplan-Meier Gráfico: Para cada tiempo exacto se incluye, según el método de Kaplan-Meier, la curva de probabilidad de supervivencia.

Opciones:

- Se identifica código para datos censurados.

- Obtener las probabilidades de supervivencia para todos los casos o estratificarlas por grupos de una variable explicativa del modelo.

- La cabecera, el título, el mínimo, máximo e incremento del eje X y del eje Y.

Modelo: Se muestra el proceso iterativo de disminución del estadístico "-2Log Likelihood", la ecuación del modelo, los "Odds ratio" y la tabla de clasificación.

Opciones:

- Se identifica código para datos censurados.

- Selección de variables: incluir todas, aplicar un procedimiento paso a paso hacia adelante o hacia detrás.

- Nivel de significación p para entrar y para salir en el proceso secuencial de selección del modelo: las variables van entrando en el modelo si realmente lo mejoran más allá de lo que podría deberse al azar (p-para-entrar) y pueden salir si no mejoran el modelo significativamente (con relación a p-para-salir).

- Valor alfa para construir los intervalos de confianza para el OR asociado a cada variable explicativa del modelo. En las opciones el valor por defecto de alfa es 5% que corresponde a un IC del 95%. Alfa debe ser >0 y <100.

- Número máximo de iteraciones que se permite.

Por defecto, el programa tiene marcadas las opciones: código de censura igual al primer valor en el fichero de datos para la variable censura, incluidas todas las variables, p-para-entrar y p-para-salir con valor 0.1, alfa de un 5% y el número máximo de iteraciones es 20.

Aunque el algoritmo utilizado es robusto en la mayoría de situaciones, no se asegura la convergencia en todas los casos.

Se desea modelizar la variable Meses mediante las variables Sexo, Antigüedad y Edad mediante un modelo de regresión de Cox utilizando Cens como variable de censura (código=0) y el resto de variables como variables explicativas. Los datos son los siguientes:

Meses	Cens	Sexo	Antigüedad	Edad
12	1	1	2	45
12	1	1	2	45
12	1	1	2	45
23	0	0	2	34
23	0	0	2	34
24	1	1	3	23
24	1	1	3	23
24	1	1	3	23
25	1	0	2	34
25	1	0	2	34
34	1	1	6	45
35	1	0	4	56
35	1	0	4	56
45	1	1	6	45
45	1	1	6	45
45	1	1	6	45
46	1	0	5	34
46	1	0	5	34
45	0	1	4	23
45	1	1	5	34
45	1	1	5	34
34	1	0	6	45

Resultados estadísticos parciales.

Regresión de Cox. Estadísticos

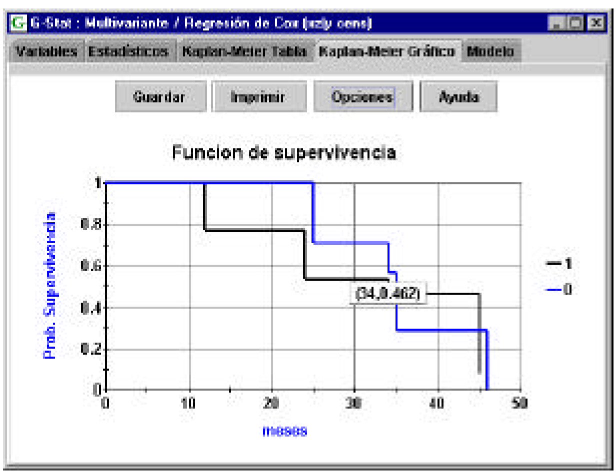
Variable Respuesta: meses
Variable de Censura: cens
Variable(s) Explicativa(s): sexo, anti, edad
Número de Casos: 22

Variable=meses

cens	N	Media	Mediana	Desviación Típica	Mínimo	Máximo
1	19	32.2632	34.0000	12.3641	12.0000	46.0000
0	3	30.3333	23.0000	12.7017	23.0000	45.0000
Total	22	32.0000	34.0000	12.1185	12.0000	46.0000

Variable=sexo

cens	N	Media	Mediana	Desviación Típica	Mínimo	Máximo
1	19	0.6316	1.0000	0.4956	0.0000	1.0000
0	3	0.3333	0.0000	0.5774	0.0000	1.0000
Total	22	0.5909	1.0000	0.5032	0.0000	1.0000



Curvas de la función de supervivencia por Kaplan-Meier para los grupos formados por la variable Sexo con código de dato censurado=0.

Resultados del Modelo de la regresión de Cox, con selección de todas las variables sin mostrar las iteraciones.

```
Regresión de Cox. Modelo
======================================================================

Variable Respuesta:                meses
Variable Censura:                  cens
Valor asociado al dato censurado:  cens=0
Valor asociado al dato exacto:     cens=1
Variable(s) Explicativa(s):        sexo, anti, edad
Número de Casos:                   22
Número de casos censurados:        3
Número de casos exactos:           19

Número de casos excluidos (*):          0
(*) Correspondientes a casos censurados antes del primer evento

 Función Logaritmo de la Verosimilitud LL (Log Likelihood)
------------------------------------------------------------
 -2 Log Likelihood = 88.8187 (Modelo Nulo)
 -2 Log Likelihood = 70.3077
 -2 Log Likelihood = 68.7764
 -2 Log Likelihood = 68.6199
 -2 Log Likelihood = 68.6141
 -2 Log Likelihood = 68.6141
 -2 Log Likelihood = 68.6141 (Modelo Final)

Ajuste del Modelo (Diferencia de Likelihood)
------------------------------------------------------------------
 Chi-Cuadrado =    20.2047;    G.L.    3;    p-valor =    0.0002

Regresión de Cox
--------------------------------------------------------------------
 Variable   Coef.      E.E.     Wald     G.L.    p-valor    R
--------------------------------------------------------------------
 sexo       2.7028    1.3683   3.9016   1       0.0482     0.1463
 anti       -1.4654   0.5059   8.3910   1       0.0038     -0.2682
 edad       0.1072    0.0494   4.7147   1       0.0299     0.1748

 Variable   HR         IC95.0%inf   IC95.0%sup
-----------------------------------------------
 sexo       14.9217   1.0211       218.0465
 anti       0.2310    0.0857       0.6226
 edad       1.1132    1.0105       1.2263
```

Menú Ayuda

Activar la opción **Ayuda** del menú principal o mediante Alt+Y. Este menú contiene las opciones relacionados con la ayuda del programa. Estas opciones conectan con la pagina www.g-stat.es, donde se encuentra actualizada dicha información.

Manual del G-Stat

Contiene información actualizada en Internet sobre este manual.

Dónde Encontrar

Contiene las rutas de los diferentes análisis y técnicas estadísticas contenidas en este programa.

Acerca de G-Stat

Contiene información actualizada sobre los créditos del programa y condiciones de utilización.

Bibliografía

Armitage, P.; Berry, G.; Matthews, J.N.S. *Statistical Methods in Medical Research.* Blackwell Science Publications, Oxford, **2002**.

Box, G.E.P.; Hunter, W.G.; Hunter, J.S. *Statistics for Experimenters.* John Wiley & Sons, New York, **1978**.

Clegg, F. *Estadística Fácil. Aplicada a las Ciencias Sociales.* Grijalbo, Barcelona, **1984**.

Collett, D. *Modelling Binary Data.* Chapman & Hall, Londres, **1991**.

Collett, D. *Modelling Survival Data in Medical Research.* Chapman & Hall, Londres, **1994**.

Daniel, W.W. *Applied Nonparametric Statistics.* PWS-KENT Publishing Company, Boston, **1990**.

Desu, M.M.; Raghavarao, D. *Nonparametric statistical methods for complete and censored data.* Chapmam & Hall/CRC, Florida, **2004**.

Fleiss, J.L. *Statistical Methods for Rates and Proportions.* John Wiley & Sons, New York, **1981**.

Gonick, L.; Smith, W.*The Cartoon Guide to Statistics.* HaperPerennial, New York, **1993**.

Hosmer, D.W; Lemeshow, S. *Applied Logistic Regression.* John Wiley & Sons, New York, **1989**.

Juez, P.; Díez, F.C. *Probabilidad y Estadística en Medicina.* Diaz de Santos, Madrid, **1996**.

Kleinbaum, D.G.; Kupper. L.L.; Muller, K.E.; Nizam, A. *Applied Regression Analysis and Multivariable Methods.* Duxbury Press, Pac. Grove, **1998**.

Lee, E.T. *Statistical Methods for Survival Data Analysis.* John Wiley & Sons, New York, **1992**.

Letón, E.; Pedromingo, A. *Fundamentos Teóricos del Análisis de Supervivencia*. GSK, Madrid, **1996**.

Letón, E.; Pedromingo, A. *Introducción al Análisis de Datos en Meta-Análisis*. Díaz de Santos, Madrid, **2001**.

Marubini, E.; Valsecchi, M.G. *Analysing Survival Data from Clinical Trials and Observational Studies.* John Wiley & Sons, Chichester, **1995**.

Matthews, D.E.; Farewell, V.T. *Using and Understanding Medical Statistics.* Karger, Basel, **1988**.

Norman, G.; Streiner, D. *Biostatistics: The Bare Essentials.* Mosby Year Book, St. Louis, **1994**.

Pedromingo, A.; Letón, E. *118 ejercicios de Estadística Básica aplicados a las Ciencias de la Salud*. GSK, Madrid, **1994**.

Ríos, S. *Iniciación Estadística.* Paraninfo, S.A., Madrid, **1992**.

Snedecor, G.W.; Corchran, W.G. *Statistical Methods.* The Iowa State University Press, Iowa, **1980**.

Wang, C. *Sense and Nonsense of Statistical Inference.* Marcel Dekker, New York, **1993**.

Wooding, W.M. *Planning Pharmaceutical Clinical Trials.* John Wiley & Sons, New York, **1994**.