InfoStat

Software Estadístico

Manual del Usuario

Versión 2008

InfoStat

Manual del Usuario

Versión 2008

El software y la documentación de InfoStat es el resultado de la participación activa y multidisciplinaria de todos los miembros del Grupo InfoStat, quienes son titulares del Copyright. La asignación de roles y actividades principales son:

Programación: **Julio A. Di Rienzo**

Algoritmos estadísticos: **Mónica G. Balzarini, Julio A. Di Rienzo, Carlos W. Robledo**

Control de calidad: **Fernando Casanoves**

Dirección editorial del Manual del Usuario: **Mónica G. Balzarini**.

Edición electrónica del Manual: **Laura A. Gonzalez**

Ayuda en línea: **Elena M. Tablada**

La cita bibliográfica correcta para este manual es como sigue:

Balzarini M.G., Gonzalez L., Tablada M., Casanoves F., Di Rienzo J.A., Robledo C.W. (2008). *Manual del Usuario*, Editorial Brujas, Córdoba, Argentina.

La obra de software a la que se refiere este manual debe citarse en bibliografía como sigue:

Di Rienzo J.A., Casanoves F., Balzarini M.G., Gonzalez L., Tablada M., Robledo C.W. (2008). *InfoStat, versión 2008,* Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina.

Prólogo

InfoStat es un software estadístico desarrollado por el *Grupo InfoStat*, un equipo de trabajo conformado por profesionales de la Estadística Aplicada con sede en la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba. Por la Cátedra de Estadística y Biometría participaron en la elaboración de InfoStat los profesores. **Julio A. Di Rienzo**, **Mónica G. Balzarini**, **Fernando Casanoves**, **Laura A. Gonzalez**, **Elena M. Tablada** y por la Cátedra de Diseño de Experimentos participó el Prof. **Carlos W. Robledo**. InfoStat, como proyecto de investigación y desarrollo representa una síntesis de la experiencia acumulada desde 1982 en la Unidad de Procesamiento Electrónico de Datos y en la Cátedras de Estadística y de Diseño de Experimentos. Labor enriquecida por la tarea docente de grado y postgrado, la consultoría estadística y la formación de recursos humanos en estadística aplicada realizada por los miembros del equipo de desarrollo. Nos enorgullece hoy el nivel de aceptación que InfoStat tiene en el ámbito universitario, en centros de investigación y tecnología y en empresas de producción de bienes y servicios.

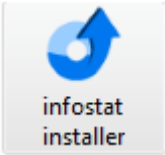
El presente manual consta de cuatro capítulos: Manejo de Datos, Estadísticas, Gráficos y Aplicaciones. El capítulo Manejo de Datos contiene información acerca de cómo operar el programa para la utilización de archivos y describe las acciones que se pueden realizar sobre tablas de datos. El capítulo Estadísticas describe las herramientas metodológicas que el usuario puede seleccionar para el análisis de sus datos. Estas descripciones están acompañadas de ejemplos de su implementación en InfoStat construido a partir de numerosas situaciones reales donde la aplicación de una o más técnicas estadísticas resulta beneficiosa. El capítulo Gráficos describe también a través de la ejemplificación diferentes tipos de representaciones gráficas disponibles. El capítulo Aplicaciones presenta métodos estadísticos usados en el control estadístico de la calidad, la cuantificación de biodiversidad y herramientas computacionales para facilitar el proceso de enseñanza – aprendizaje de conceptos clásicos de la estadística.

El presente manual refleja el estado de avance de InfoStat al momento de su impresión, sin embargo InfoStat está creciendo, actualizando y mejorando continuamente algoritmos e interfases con el usuario. En el menú ayuda de InfoStat encontrará acceso a la versión a la versión electrónica de este manual y a un link para su actualización en línea.

Índice de contenidos

Instalación

Para la instalación de InfoStat desde la página web www.infostat.com.ar se deben seguir las instrucciones que allí se detallan. Para la instalación desde un CD de distribución, la computadora requiere unos segundos para leer los archivos de iniciación. Este proceso puede demorar hasta 2 minutos en algunas máquinas. Una vez que se inicia el proceso de instalación sólo apriete la tecla <Enter> en cada ventana de opciones que se le presente en pantalla. Cuando el proceso concluye exitosamente entonces el instalador habrá creado una carpeta InfoStat dentro de C:\Archivos de programa\ y un ícono de acceso directo en Inicio→Programas→InfoStat. Si eventualmente el CD no inicia el proceso de instalación automática entonces abra el directorio del mismo, busque el ícono que tiene como leyenda **InfoStatInstaller.exe** y haga doble *click* sobre él.



Dentro de la carpeta de InfoStat, C:\Archivos de Programa\InfoStat, se encontrará la siguiente información:

Carpeta **Datos**: contiene todos los archivos de datos a los que hace referencia este manual.

Carpeta **Ayuda**: contiene el archivo de ayuda en línea.

Archivo **Manual.pdf**: contiene el material impreso que se recibió junto con el C.D. La versión electrónica del manual puede contener actualizaciones de este material impreso.

Actualización

Puede acceder a las instrucciones de actualización a través del menú AYUDA. La opción ACTUALIZAR abre la página web de InfoStat desde donde puede bajar las últimas actualizaciones.

Requerimientos

Procesador requerido: Tipo Pentium o superior

Memoria mínima sugerida: 128 Mb

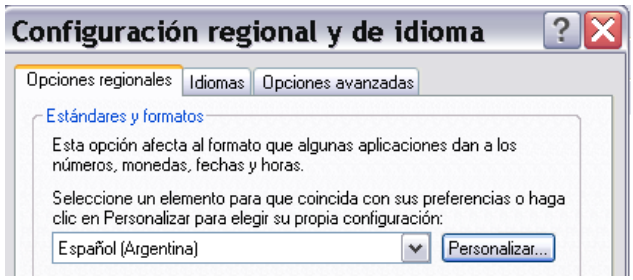
Sistemas operativos: Windows 98, 2000, XP, NT, Vista.

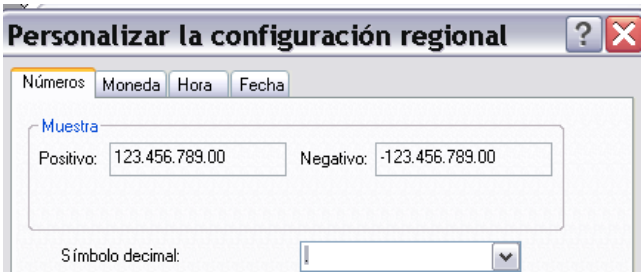
Configuración del monitor: definición mínima 800x600 píxeles, fuentes pequeñas. Si las fuentes de la configuración del monitor son grandes se pueden presentar problemas para ver parte de las ventanas que InfoStat despliega durante el trabajo. Bajo Windows 98 las fuentes pequeñas se especifican pidiendo **Propiedades** del monitor y seleccionando la solapa **Configuración.** Allí apretar el botón **Avanzada**, lo que conducirá a una ventana de diálogo en la que se puede especificar el tamaño de las tipografías del monitor.

IMPORTANTE: InfoStat reconoce automáticamente la configuración regional de la computadora. Esta, define entre otras cosas, el símbolo a utilizar como separador decimal, ya sea coma "**,**" o punto "**.**". Por defecto, las versiones en español de Windows configuran su sistema para que reconozca a la coma como separador decimal. Si la computadora está configurada para reconocer comas, cuando se use punto como separador decimal durante el ingreso de datos desde el teclado, InfoStat considerará que se ingresó un conjunto de caracteres alfanuméricos y no un valor numérico y por lo tanto no podrá tratarlos para hacer cálculos. Este comportamiento es común a todas las aplicaciones Windows, pero se hace crítico cuando la aplicación procesa datos numéricos cargados por el usuario.


Configuración regional

Si Ud. quiere cambiar la configuración regional para utilizar punto (o coma), debe entrar a **Panel de Control** (Menú Inicio→Configuración→Panel de Control) y localizar el ícono de la configuración regional. Haga doble *click* sobre ese ícono y aparecerá una ventana cuya parte superior se muestra a continuación. Una opción es simplemente cambiar la configuración regional eligiendo Estados Unidos, por ejemplo. La otra es tocando la solapa **Número** y cambiando allí el símbolo decimal. Esta opción es probablemente más recomendable ya que las otras pueden cambiar los estilos de fechas y criterios de ordenamiento alfabético.





Aspectos generales

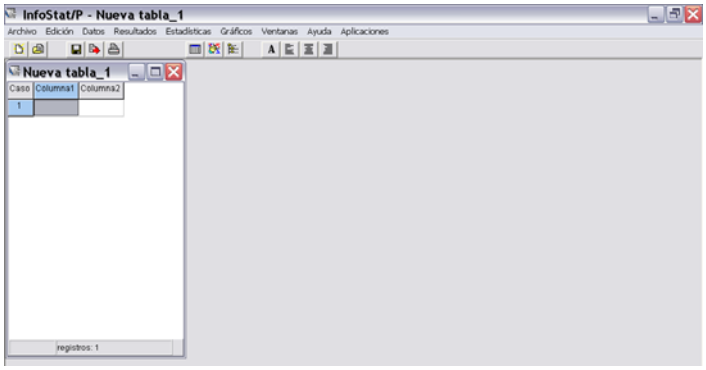
InfoStat ofrece distintas herramientas para que el usuario pueda explorar su información de manera muy sencilla. Al abrir InfoStat, se visualizará una barra de herramientas localizada en la parte superior de la ventana del programa, la que contiene los siguientes menués: **Archivo**, **Edición**, **Datos**, **Resultados**, **Estadísticas**, **Gráficos**, **Ventanas**, **Ayuda** y **Aplicaciones**.

Por debajo de los menúes, la barra de herramientas contiene una serie de botones que permiten invocar acciones de manera rápida. Todas las acciones que se llevan a cabo con los botones también pueden ser realizadas a partir de alguno de los menúes listados arriba.

Posicionándose sobre un botón, sin presionar el ratón, el usuario visualizará una etiqueta de ayuda sobre el botón y una leyenda al pie de la pantalla indicando el tipo de acción que puede realizar con ese botón. Estas son (para los botones ordenados de izquierda a derecha) las siguientes: **Nueva tabla**, **Abrir tabla**, **Guardar tabla**, **Exportar Tabla**, **Imprimir**, **Agregar nueva columna**, **Ordenar**, **Editar Categorías**, **Fuente**, **Alineación a izquierda**, **Alineación al centro** y **Alineación a derecha**.

Al pie de la pantalla el usuario visualizará tres ventanas minimizadas, una denominada **Resultados**, otra **Gráficos** y otra **Herramientas gráficas**. Si se maximiza la ventana **Resultados** cuando recién se abre el programa, InfoStat reportará que no hay resultados disponibles. Esta ventana irá recibiendo contenido a medida que se ejecuten acciones (análisis) que produzcan resultados. Las ventanas **Gráficos** y **Herramientas Gráficas** sólo se activan cuando se ha producido un gráfico.

En el menú ARCHIVO InfoStat permite abrir y guardar archivos de datos de distintos tipos. Por ejemplo, si se acciona **Nueva Tabla** se visualizará la siguiente pantalla:



El usuario podrá ingresar información, desde el teclado, en la tabla o archivo denominado temporariamente como *Nueva*. Sobre esa tabla podrá realizar análisis de datos y producir resultados y gráficos. En el menú ARCHIVO también se encuentra el comando **Salir** para cerrar la aplicación.

En el menú EDICIÓN se encuentran los comandos para cortar, copiar y pegar información desde ventanas de datos, resultados y gráficos. El menú DATOS permite realizar operaciones de diversa índole sobre la grilla de datos; entre otras, es posible ordenar el archivo, transformar columnas, generar nuevas columnas a partir de fórmulas, simular realizaciones de variables aleatorias, buscar y reemplazar información de manera automática. Desde el menú RESULTADOS se pueden invocar acciones relacionadas a la presentación y a la exportación de resultados en formato de tabla.

Todos los resultados producidos (tablas y gráficos) pueden ser copiados utilizando el menú EDICIÓN (**Copiar**) y luego pegados en el procesador de texto, siendo ésta la manera más simple de transportar los resultados de InfoStat a un documento o informe escrito. El uso de los comandos **Copiar** y **Pegar** también es la forma más sencilla de importar y exportar datos entre InfoStat y un procesador de texto o una planilla electrónica como por ejemplo Excel. Para simplificar la migración de planillas de datos, InfoStat provee al usuario con los comandos **Copiar** y **Pegar con nombres de columnas** para conservar los nombres o etiquetas de columnas. También es posible importar y exportar información en formato ASCII. Las opciones de los menúes ARCHIVO, EDICIÓN, DATOS y RESULTADOS se detallan y ejemplifican en este capítulo.

InfoStat trabaja con tres tipos de ventanas: la ventana donde se encuentran los datos (**Datos**), aquella donde se muestran y acumulan los resultados de los procedimientos solicitados (**Resultados**) y la ventana donde se muestran y acumulan los gráficos realizados por el usuario (**Gráficos**). Varias ventanas de datos pueden mantenerse abiertas simultáneamente. En tal caso la ventana activa es aquella que presenta el marco superior coloreado (no gris). Todas las acciones serán ejecutadas sobre la ventana de datos activa. Las ventanas **Resultados** y **Gráficos** contienen una hoja para cada resultado y/o gráfico producido. El usuario puede moverse a través de las distintas hojas haciendo un *click* sobre las solapas que se encuentran al pie de la ventana y que indexan las salidas.

En el menú ESTADÍSTICAS InfoStat ofrece la posibilidad de implementar de manera casi automática (a través de ventanas de diálogo) una amplia serie de análisis estadísticos. El usuario podrá realizar estadística descriptiva, calcular probabilidades, estimar características poblacionales bajo distintos planes de muestreo, estadística inferencial para una y dos muestras mediante diversos tipos de intervalos de confianza y pruebas de hipótesis (paramétrica y no paramétrica), utilizar modelos de regresión y análisis de varianza para distintos tipos de experimentos diseñados y estudios observacionales, estadística inferencial para datos categorizados, estadística multivariada, análisis de series de tiempo, suavizados y ajustes.

Después de seleccionar la aplicación estadística que se desea utilizar para analizar los datos de un archivo abierto (tabla activa), se presenta una ventana (Selector de Variables) donde a la izquierda se listan todas las columnas del archivo para que el usuario seleccione la o las columnas que participarán en el análisis, ya sea como variable de interés o como criterio de clasificación. Las columnas seleccionadas deberán transportarse a la lista de **Variables** que se encuentra a la derecha de la ventana utilizando el botón que contiene la flecha "→". Si una variable fue seleccionada equivocadamente o ya no es necesaria puede eliminarse de la lista de variables y agregarse nuevamente a la lista de columnas del archivo oprimiendo la tecla "←" después de seleccionar la variable o haciendo doble *click* sobre la misma.

El selector de variables facilita el trabajo ya que no se deben recordar ni escribir los nombres de las variables cada vez que se quieren utilizar.

En el menú GRÁFICOS, InfoStat brinda herramientas gráficas de índole profesional para la presentación de resultados. Las técnicas gráficas implementadas son variadas y se encuentran documentadas en el capítulo Gráficos. El graficador permite incluir en un mismo gráfico varias series y editar virtualmente todos sus atributos a través de la ventana **Herramientas Gráficas** que se abre automáticamente al solicitar un gráfico. InfoStat cuenta con un algoritmo de copia y suscripción de formato que facilita la creación de series de gráficos con idénticas características. Los gráficos creados por InfoStat pueden ser guardados o copiados y pegados a cualquier aplicación Windows que soporte imágenes (metarchivo mejorado) usando los clásicos comandos Windows de copiado y pegado (o pegado especial). Todas las herramientas del menú GRÁFICOS se encuentran disponibles en todas las versiones de InfoStat.

A través del menú VENTANAS el usuario puede migrar de una ventana a otra. Otra forma de acceder a una ventana, es simplemente moviendo el cursor hacia la ventana deseada. El menú ventanas también permite seleccionar el modo en que las ventanas abiertas serán presentadas en pantalla. Estas pueden estar en cascada, presentación vertical u horizontal según el usuario haga un *click* sobre la opción **Cascada**, **Ordenar vertical** u **Ordenar horizontal**, respectivamente. A partir de este menú se puede acceder a la ventana Resultados, donde se acumulan los resultados de una sesión que el usuario no haya borrado deliberadamente. De la misma manera se puede migrar a la ventana Gráficos. Además se listan los nombres de las tablas de datos abiertas.

En el menú AYUDA se puede acceder a documentación en línea sobre procedimientos y análisis estadísticos posibles de implementar desde cualquier menú habilitado y al manual de InfoStat en formato electrónico. Además es posible usar este menú para tener un acceso rápido a la actualización del software.

Bajo el menú APLICACIONES se presentan herramientas de análisis tradicionales que son utilizadas para la exploración de información en conjuntos de datos provenientes de áreas específicas del conocimiento. Las aplicaciones disponibles son: CONTROL DE CALIDAD, DIDÁCTICAS, ÏNDICES y MICROMATRICES DE ADN. La aplicación DIDÁCTICAS está orientada a brindar elementos clásicos para la enseñanza y el aprendizaje de la estadística aplicada. Algunas herramientas frecuentemente usadas en el control estadístico de calidad, se encuentran en CONTROL DE CALIDAD. Bajo el ítem ÏNDICES, el usuario puede calcular numerosos índices de biodiversidad comúnmente usados en Ecología. En MICROMATRICES DE ADN están disponibles procedimientos de normalización, transformación, filtrado, agrupación y ordenación de genes, ordenación de micromatrices, corrección de p valor para controlar tasa de descubrimientos falsos (FDR), test de hipótesis, entre otros.

Cuando una opción de cualquier menú se presenta en color gris en vez de negro significa que la misma no está habilitada. Puede suceder que el usuario no haya cumplimentado un paso previo requerido para esa acción o que la misma no se encuentre disponible en la versión de InfoStat adquirida.

Manejo de datos

InfoStat procesa la información proveniente de una tabla. Una tabla se define como un agrupamiento de datos dispuestos en filas y columnas. Las columnas representan usualmente a las **variables** y las filas a las **observaciones**. Las etiquetas de las columnas son los nombres con que se designan las variables.

Archivo

Las acciones (submenúes) que se aplican al manejo de tablas en el menú ARCHIVO son:

NUEVA TABLA, ABRIR…, GUARDAR TABLA, GUARDAR TABLA COMO... y CERRAR TABLA. También en esta ventana se dispone de la opción SALIR y de una lista de los últimos archivos trabajados.

Nueva tabla

Menú ARCHIVO ⇒ NUEVA TABLA, crea una nueva tabla. También puede presionar <Ctrl+N> o usar el botón con la hoja en blanco de la barra de herramientas (botón **Nueva Tabla**). Aparecerá una tabla con una fila y dos columnas que podrá ampliarse para ingresar sus datos. Las tablas nuevas tienen en su nombre numeración consecutiva (Nueva tabla, Nueva tabla _1, Nueva tabla_2, etc.).

Abrir tabla

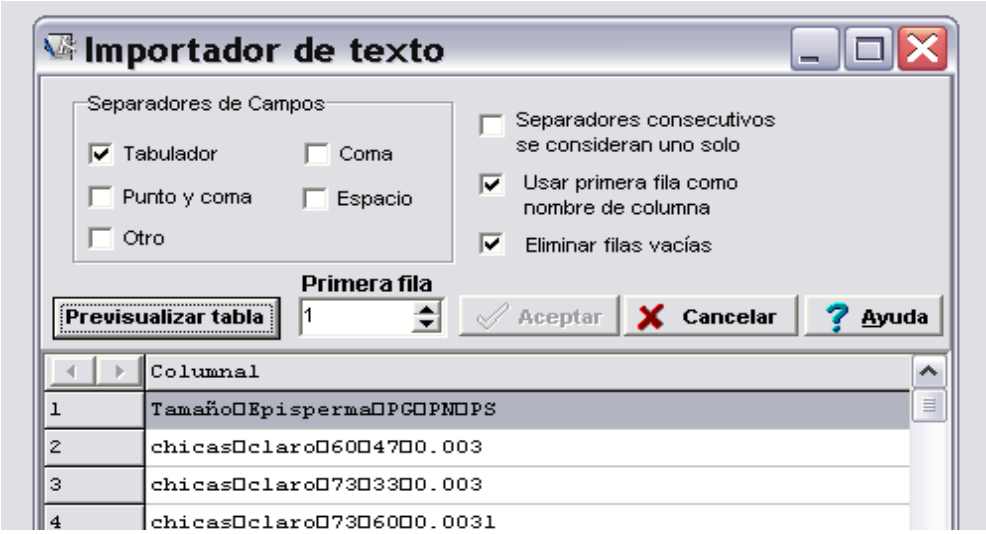
Menú ARCHIVO ⇒ ABRIR …, invoca una tabla existente. También puede presionar <Ctrl+O> o usar el botón con el dibujo de una carpeta (botón **Abrir Tabla**), de la barra de herramientas. Activando <Shift>+botón **Abrir Tabla** se accede directamente a la carpeta Datos la cual contiene los archivos de los ejemplos dados en este manual. Para abrir una tabla, en la ventana de diálogo complete la información solicitada.

InfoStat permite abrir archivos con los siguientes formatos:

InfoStat (*.IDB, *.IDB2)	Excel (*.XLS)	Gráficos (*.IGB)
Textos (*.TXT, *.DAT)	Dbase (*.DBF)	Resultados (*.ITRES)
InfoGen (*.IGDB)	Paradox (*.DB)	EpiInfo (*.REC)

InfoStat asume que en la estructura de los datos las columnas representan a las variables y las filas a las observaciones. Para cada variable todos los valores deberán corresponder al mismo tipo de dato (entero, real, categoría o fecha).

Si desea abrir un archivo ASCII, con extensión TXT o DAT, se activará la ventana del **Importador de texto**.



Con el **Importador de texto** se podrá indicar: el o los caracteres **Separadores de campos** que desea utilizar (tabulador, coma, punto y coma, espacio u otros). Los datos a importar pueden contener o no el nombre de las variables (columnas). Si los datos contienen el nombre de las columnas, se puede indicar si lo que aparece en la **Primera fila** será el nombre de las futuras columnas de la tabla de datos (InfoStat muestra esta opción por defecto). Si en el encabezado figura algún texto antes de los nombres de las columnas, se deberá indicar qué línea contiene el nombre de las columnas; esto se hace cambiando el número que está al lado de la opción **Primera fila**, hasta que se visualice la línea con el nombre de las columnas en primera fila. Si los datos no contienen nombre de columnas, se deberá deseleccionar la opción **Usar primera fila como nombre de columna**. En este caso las variables serán encabezadas como Columna1, Columna2, etc. Para observar la información, que constituirá la tabla una vez importada, presionar el botón **Previsualizar Tabla**. Si la estructura es correcta presionar **Aceptar**, caso contrario, cambiar opciones y probar nuevamente con **Previsualizar tabla** hasta obtener el resultado deseado.

Si al previsualizar la tabla se observa que los nombres de las columnas aparecen desplazados respecto del contenido de la tabla, se pueden acomodar los nombres utilizando los botones de desplazamiento que se encuentran en la parte superior de la columna que identifica a las filas de la tabla que se previsualiza.

Nota: cuando se importan tablas de archivos Microsoft Excel que han sido grabados como texto (con extensión .TXT), las celdas vacías en el archivo original se muestran como dos separadores consecutivos en el archivo de texto, en tal caso la opción **Separadores consecutivos se consideran uno solo**, no debe ser seleccionada. Por defecto, InfoStat muestra esta opción no seleccionada cuando abre archivos de texto. Por otra parte si el archivo contiene datos numéricos y alfanuméricos, en una misma columna, InfoStat sólo reconocerá el primer carácter de la columna, si es un número borrará los alfanuméricos de la columna y viceversa. La forma más sencilla de leer archivos desde otro programa es con las funciones Copiar y Pegar. InfoStat ofrece las opciones **Copiar con nombre de columna** y **Pegar con nombre de columna** para facilitar la importación y exportación de datos. Por ejemplo, para importar un archivo Excel simplemente copie los datos que desea llevar a InfoStat incluyendo el nombre de las columnas desde Excel y abra una nueva tabla en InfoStat donde deberá pegar usando la opción **Pegar con nombre de columna** el contenido copiado.

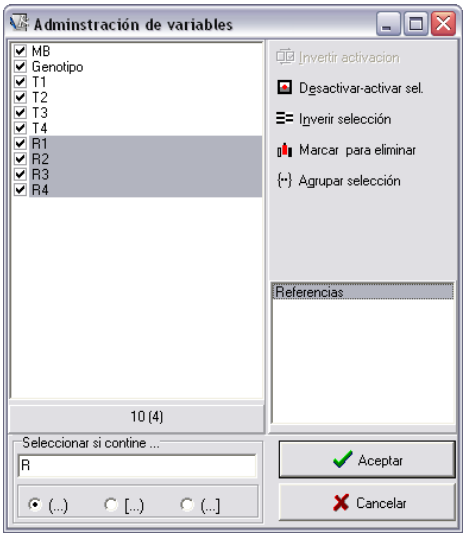
Barra de herramientas de la tabla

Al ubicar el cursor sobre una tabla, si se presiona el botón derecho del ratón se dispone de varias opciones entre las cuales se halla **Barra de herramientas**. Con esta opción se agrega, a la tabla activa, una barra de botones:

Estos botones, de izquierda a derecha, permiten: aumentar el tamaño de la fuente, disminuir el tamaño de la fuente, quitar decimales (previamente se debe hacer clic en una celda de la columna de interés), agregar decimales (previamente se debe hacer clic en una celda de la columna de interés), insertar una fila (antes de una fila previamente seleccionada), eliminar una fila previamente seleccionada, agregar una columna al final de la tabla, insertar una columna (antes de una columna previamente seleccionada), eliminar una columna previamente seleccionada y colorear una selección.

El tamaño de la fuente también puede ser modificado si se presionan las teclas Ctrl. y ↑ (para agrandar), o Ctrl. y ↓ (para disminuir).

Administración de variables



Esta ventana aparece cuando se tiene una tabla activa y se presionan las teclas <Ctrl+E>. El conjunto de acciones disponibles en el diálogo son:

Renombrar variables: haciendo doble clic sobre un nombre de la lista de variables, se puede modificar el mismo.

Mover la posición de una o más variables: de la lista se seleccionan las variables y teniendo presionada la tecla Ctrl., se mueve el bloque seleccionado utilizando las teclas de dirección (↑ mueve hacia arriba y ↓ mueve hacia abajo). Los

cambios de posición realizados en la lista se actualizan automáticamente en la tabla.

Marcar una o más variables para eliminar: se seleccionan las variables de la lista y se hace clic en el botón **Marcar para eliminar**. Las variables son eliminadas de la lista y de la tabla.

Desactivar- activar una o mas variables: La condición de desactivada se indica cuando el cuadro de chequeo a la izquierda de la etiqueta aparece sin el tilde (en el ejemplo están todas activadas y seleccionadas todas las que contienen un 1 en la etiqueta). Las variables desactivadas no aparecen en la tabla ni en el selector de variables).

Formar grupos de variables: seleccionando variables y apretando el botón **Agrupar selección,** se pueden formar grupos de variables que luego pueden activarse-desactivarse conjuntamente, colorearse, borrarse, etc.

Guardar tabla

Menú ARCHIVO ⇒ GUARDAR TABLA, guarda la tabla activa en formato InfoStat (con extensión **.IDB2)**, en el directorio en uso. También puede presionar <Ctrl+S>, o el botón **Guardar Tabla** de la barra de herramientas.

Guardar tabla como

Menú ARCHIVO ⇒ GUARDAR TABLA COMO, guarda la tabla activa con el formato y en el directorio requerido por el usuario. Los formatos son:

InfoStat (*.IDB, *. IDB2)	Excel (*.XLS)	Dbase (*.DBF)
ASCII (*.TXT)	InfoGen (*.IGDB)	Paradox (*.DB)

También se puede utilizar el botón **Exportar Tabla** de la barra de herramientas.

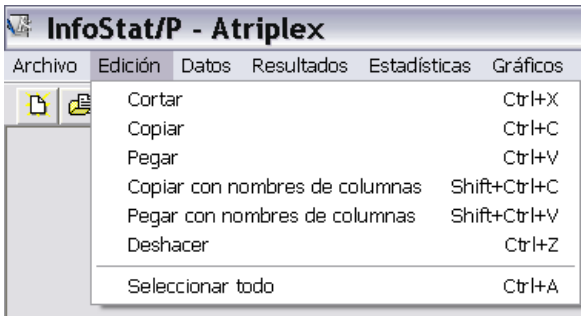
En la ventana de diálogo indique el nombre, lugar y el tipo de archivo. Si elige el formato ASCII deberá seleccionar el separador de campos, indicar si desea usar la primera fila como nombre de columnas (etiquetas) y opcionalmente indicar un carácter (o grupo de caracteres) para identificar una observación faltante en el archivo exportado.

Cerrar tabla

Menú ARCHIVO ⇒ CERRAR TABLA cierra la tabla activa. También se puede presionar <Ctrl+W>. Si la tabla ha sido modificada y no ha sido guardada, InfoStat le pedirá que confirme si desea guardarla.

Edición

Las acciones (submenúes) que se aplican al manejo de tablas InfoStat en el menú EDICIÓN son: **Cortar, Copiar, Pegar, Copiar con nombre de columna, Pegar con nombre de columna, Deshacer** y **Seleccionar todo**. Las acciones se utilizan para edición de celdas, columnas y/o filas, como es usual en edición de textos bajo Windows.

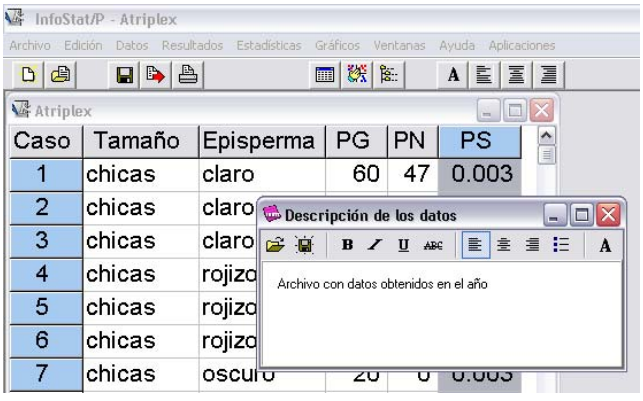


Las modificaciones de datos ingresados en una tabla de InfoStat se hacen en la celda activa. Presione <Enter> para que los caracteres ingresados sean cargados en la tabla. Si antes de presionar <Enter> se presiona la tecla <Esc>, se establecerá de nuevo el contenido anteriormente cargado en la celda. Para salir de la edición de la celda use las teclas de direccionamiento (las flechas arriba, abajo, izquierda o derecha), tabulador o seleccione con el ratón otra celda.

Para seleccionar un conjunto de celdas seleccione con el ratón el área deseada. También puede seleccionar celdas con el teclado manteniendo presionada la tecla mayúscula (<Shift>) y moviéndose con las teclas de dirección. Las áreas pintadas pueden ser impresas apretando el botón de **Impresión** de la barra de herramientas.

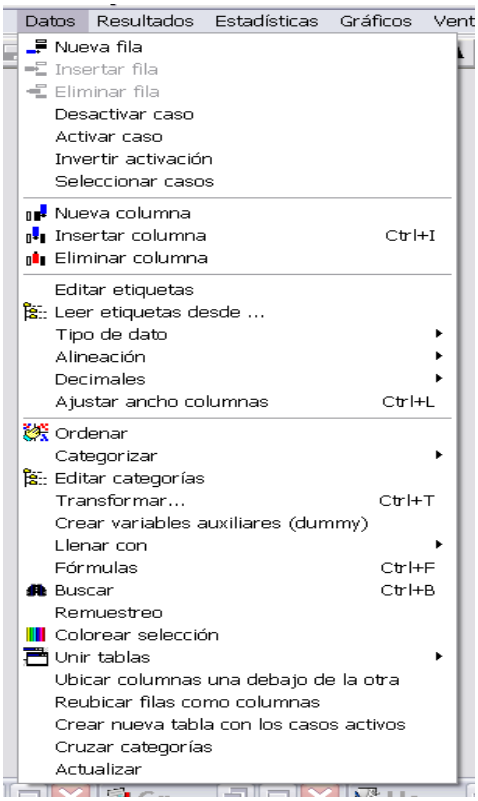
Es posible elegir el tipo, estilo, tamaño y color de letra en toda la tabla, sólo basta seleccionar una celda y presionar el botón con el carácter "**A**" de la barra de herramientas para obtener el menú apropiado para realizar esta acción. También existen botones para alineación derecha, izquierda y al centro de la columna de datos. Dichos botones se encuentran al lado del botón "**A**".

En tablas con formato .IDB2 se puede guardar una descripción acerca de los datos que contiene la tabla. La descripción se edita presionando F2. Aparece un campo en el que se escribe la descripción, la cual será incrustada en el archivo si se presiona el segundo botón de la barra de herramientas de la ventana de diálogo. Para incorporar definitivamente la descripción al archivo de datos, se debe guardar la tabla.



Una descripción puede ser cargada desde un archivo con formato txt o rtf, si se presiona el primer botón de la mencionada barra de herramientas.

Datos



Las acciones (submenúes) que se aplican al manejo de tablas InfoStat en el menú DATOS son: **Nueva fila**, **Insertar fila**, **Eliminar fila**, **Desactivar caso**, **Activar caso**, **Invertir activación, Seleccionar casos, Nueva columna**, **Insertar columna**, **Eliminar columna**, **Editar Etiquetas**, **Leer etiquetas desde…, Tipo de dato**, **Alineación**, **Decimales**, **Ajustar ancho columna**, **Ordenar**, **Categorizar**, **Editar categorías**, **Transformar …**, **Crear variables auxiliares (dummy)**, **Llenar con...**, **Fórmulas**, **Buscar**, **Remuestreo**, **Colorear selección**, **Unir tablas**, **Ubicar columnas una debajo de la otra**, **Reubicar filas como columnas**, **Crear nueva tabla con los casos activos, Cruzar Categorías** y **Actualizar**.

Estas acciones también pueden ser invocadas presionando el botón derecho del ratón, cuando se está posicionado en la tabla de datos.

Para ilustrar algunas de las acciones ejecutadas por los submenúes se utilizará el siguiente ejemplo:

Ejemplo 1: se dispone de un conjunto de observaciones que hacen referencia al tamaño de la semilla (Tamaño), color del episperma (Episperma), porcentaje de germinación (PG), número de plántulas normales (PN) y peso seco (PS) de semillas de Atriplex cordobensis, un arbusto forrajero. Los datos se encuentran en el archivo Atriplex.idb (gentileza Dra, M.T. Aiazzi, Facultad de Ciencias Agropecuarias, U.N.C.).

Nota: en C:\Archivos de Programa\InfoStat\Datos, se encuentran los archivos utilizados en el presente manual.

Nueva fila

Menú DATOS ⇒ NUEVA FILA, agrega al final de la tabla la cantidad de filas que especifique el usuario en la ventana emergente. También puede posicionarse en la última fila y presionar la tecla <Enter> para generar nuevas filas.

Insertar fila

Menú DATOS ⇒ INSERTAR FILA, inserta una fila por encima de la fila seleccionada.

Eliminar fila

Menú DATOS ⇒ ELIMINAR FILAS, elimina la fila o filas de la tabla que se encuentran seleccionadas. Esta acción se puede revertir usando el submenú **Deshacer** del menú **Edición**.

Desactivar caso

Menú DATOS ⇒ DESACTIVAR CASO, permite excluir del procedimiento a ejecutar las filas que se seleccionen. Para desactivar una fila de la tabla basta hacer doble clic sobre su número de caso. Las observaciones desactivadas muestran el número de caso entre paréntesis y la fila está coloreada.

Activar caso

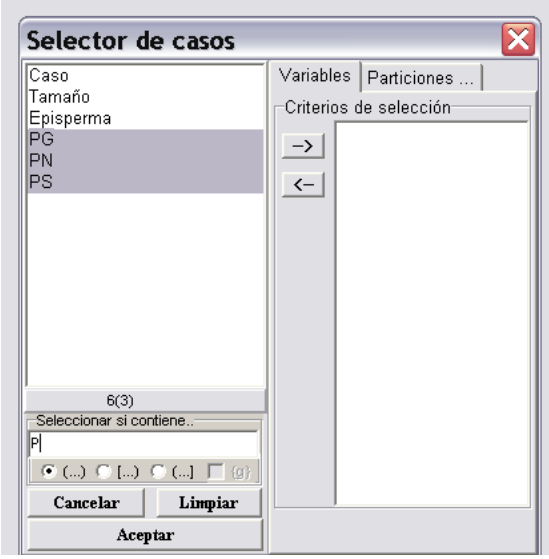
Menú DATOS ⇒ ACTIVAR CASO, hace activos (participan en el análisis) *casos* que se encuentran desactivados. Para activar una única fila basta hacer doble clic en su número de caso. Si se quieren activar varios *casos* simultáneamente basta seleccionar alguna celda de cada una de las filas a activar y aplicar esta acción desde el menú DATOS o desde el menú que aparece al presionar el botón derecho del ratón. Por defecto, todos los casos se encuentran activados.

Invertir activación

Menú DATOS ⇒ INVERTIR ACTIVACIÓN, vuelve activos (desactivados) los casos que se encuentren desactivados (activados).

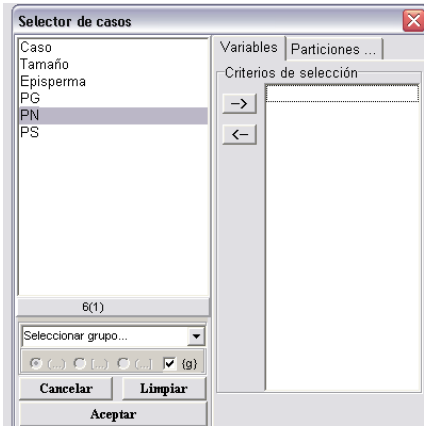
Seleccionar caso

Menú DATOS ⇒ SELECCIONAR CASO permite establecer un criterio para la selección de casos. Ejecutada la acción, los casos no seleccionados se muestran desactivados. Primero hay que establecer sobre qué variables se aplicará el criterio de selección y luego especificar el criterio.



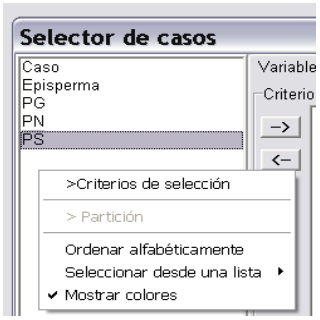
En la ventana de diálogo del **Selector de casos**, aparece la lista de las variables de la tabla activa. De dicha lista se eligen las variables sobre las que se aplicará la selección de casos, ingresándolas en el correspondiente cuadro de la **solapa Variables** (se puede indicar una partición en la correspondiente solapa).

Si se trabaja con muchas variables se dispone de procedimientos que agilizan la elección de las mismas. Al pie de la lista de variables hay opciones para elegirlas según alguna característica común de sus nombres. Si las variables comparten algún carácter o sucesión de caracteres, se pueden elegir simultáneamente. En la figura se ilustra la elección de todas las variables cuyos nombres contienen la letra P, ya que se activó la casilla de la **opción (…)**. Para especificar que el carácter o sucesión de caracteres está al inicio de la etiqueta se activa la **opción […);** para indicar que está al final de la etiqueta se activa la **opción (…]**. Se pueden usar caracteres tipo comodín. Por ejemplo, al ingresar la secuencia "****1**" quedarán seleccionadas de la lista todas las variables cuyas etiquetas tengan 2 caracteres antes del 1. Si se especifica "**??1**" se seleccionarán todas las variables cuyas etiquetas contienen un 1 precedido de dos caracteres alfabéticos y si se especifica "**##1**", serán seleccionadas todas las variables cuyas etiquetas contienen un 1 precedido de dos caracteres numéricos.



Si se han formado grupos (usando la ventana de **Administración de variables**), estará disponible la casilla con el rótulo **{g}**. Al activar esta casilla aparece un campo que contiene la lista de los grupos disponibles, permitiendo la selección de los mismos.

Otra forma de elegir variables es utilizando una lista almacenada en un archivo de texto. De esta forma se seleccionarán todas las variables contenidas en dicho archivo. Para ello, se debe presionar el botón derecho del ratón ubicando el cursor sobre el cuadro que contiene a las variables de la tabla activa. Se despliega un menú donde se elige la **opción Seleccionar desde una lista** y a continuación la **opción Archivo de texto**. En este mismo menú hay una opción para ordenar la lista de variables en forma alfabética.



Una vez elegidas las variables, se establecen los criterios para seleccionar los casos. En la ventana de diálogo aparecen las variables que intervienen en el procedimiento de selección y hay un campo para escribir el criterio. En caso de que este criterio se establezca en base a más de una variable, Se selecciona una de las variables, se escribe la sentencia que indica el criterio, por ejemplo *x<80*, y se presiona **Enter**. Luego, se procede de igual forma para cada variable de interés. Al presionar **Aceptar**, los casos fuera de la selección aparecen desactivados (coloreados y con el número de caso entre paréntesis), en la tabla activa

Se puede escribir más de una sentencia para determinar el criterio aplicado a una variable, presionando **Enter** luego de escribir la sentencia.

Además, activando la casilla **Generar nueva tabla**, se obtiene una tabla con los casos seleccionados.

Nueva columna

Menú DATOS ⇒ NUEVA COLUMNA, agrega una columna al final de la tabla. Se podrá indicar el tipo de dato (entero, real, categórico o fecha). La columna agregada recibirá el nombre de Columna1, Columna2, etc., según corresponda. Apretando el botón en la barra de herramientas que tiene el dibujo de una tabla, se agregan nuevas columnas al margen derecho de la tabla activa. Las columnas así generadas no tienen tipo asignado previamente. El tipo de dato de estas columnas se asignará automáticamente cuando se cargue un contenido en cualquiera de sus celdas. Si el contenido es numérico el tipo asignado es *real*, si es alfanumérico el tipo asignado es *categórico.* Si Ud. quiere que el tipo sea *entero* deberá cambiarlo a posteriori, partiendo de una columna con tipo *real*.

Insertar columna

Menú DATOS ⇒ INSERTAR COLUMNA, inserta una columna en la posición anterior a la que se encuentra el cursor. Se puede indicar el tipo de dato (*real, entero, categórico o fecha*), que se va a ingresar. Las columnas insertadas tendrán el nombre de Columna1, Columna2, etc.

Eliminar columna

Menú DATOS ⇒ ELIMINAR COLUMNA, elimina las columnas seleccionadas. Sólo basta con seleccionar una celda por cada columna. Esta acción se puede revertir usando el submenú **Deshacer** del menú **Edición**.

Nota: para cambiar la posición de una columna, seleccione la columna y manteniendo apretada la tecla <Ctrl> mueva el ratón manteniendo el botón apretado hasta la nueva posición. Al liberar el botón del ratón la columna quedará cambiada de lugar.

Editar Etiquetas

Menú DATOS ⇒ EDITAR ETIQUETAS, permite cambiar el nombre de una columna. Basta estar posicionado en una celda de la columna a la que se le quiere cambiar el nombre y solicitar esa acción. Los nombres permitidos aceptan espacios y caracteres ASCII, en número no mayor a veinte. Si el nombre comienza con un número InfoStat antepondrá la letra C. Seleccionando varias columnas y aplicando esta acción, se ingresa en una ventana de diálogo que permite cambiar los nombres de las columnas de manera sucesiva. Otra opción es hacer doble *click* sobre el nombre que se quiere cambiar.

En archivos generados con extensión IDB2, haciendo doble clic en el campo de edición en el que se escribe el nombre de la variable aparece un diálogo que permite introducir un texto que describa a la variable. Para incorporar la descripción al archivo se debe guardar el mismo.

Leer etiquetas desde…

Menú DATOS ⇒ LEER ETIQUETAS DESDE…, permite leer los nombres de las variables para una tabla activa, desde un archivo de texto (*.txt). InfoStat supone que los nombres están en una lista (un nombre debajo del otro), en el orden en que se encuentran las variables en la tabla.

Tipo de dato

Menú DATOS ⇒ TIPO DE DATO, permite declarar el tipo de dato contenido en una columna. Los tipos permitidos son: *entero, real, categoría* y *fecha*. Para ingresar una fecha se puede tipear de la forma 20/05/07, 20-05-07 o bien 20.5.07.

Si el tipo de dato no es declarado por el usuario, se asume como tipo el correspondiente al primer dato ingresado. Una vez que el tipo ha sido declarado, **no** se podrán ingresar datos de otro tipo.

Alineación

Menú DATOS ⇒ ALINEACIÓN, cambia la posición de la presentación de los contenidos de las celdas seleccionadas. Las posiciones de alineación son *izquierda, centro* y *derecha*. Por defecto la alineación es a la derecha para los campos numéricos e izquierda para los campos categóricos. También existen botones para realizar esta acción los que se encuentran en la barra de herramientas al lado del botón "**A**".

Decimales

Menú DATOS ⇒ DECIMALES, cambia la cantidad de decimales con los que se presentan los contenidos numéricos de las celdas. Hasta 10 decimales son permitidos. Por defecto el número de decimales es 2. En los procesos de copiado de datos desde la grilla **sólo se tienen en cuenta los decimales visibles**, por lo que es relevante especificar para cada variable el número apropiado de decimales con el que se desea trabajar.

Ajuste automático de columnas

Menú DATOS ⇒ AJUSTAR ANCHO COLUMNAS (<Ctrl+L>), ajusta el ancho de las columnas seleccionadas de acuerdo a la longitud de las etiquetas de las columnas o del contenido de las celdas. Si no se selecciona ninguna columna la acción se realiza sobre todas las columnas de la tabla.

Ordenar

Menú DATOS ⇒ ORDENAR, permite ordenar los registros en forma ascendente o descendente según los valores de una o más columnas. Una ventana de diálogo muestra en una lista a la izquierda los nombres de las columnas de la tabla activa. A la derecha aparecerán dos listas para mostrar las variables que van a ser utilizadas en el proceso de

ordenamiento según el nivel jerárquico determinado por el usuario, de acuerdo al orden en que se seleccionen las variables, y si se trata de orden ascendente o descendente. Por ejemplo si el archivo tiene dos columnas en una de las cuales figura el sexo y en la otra la edad, ubicando primero la variable sexo en la lista según criterio ascendente y en segundo lugar a la edad en la lista según criterio descendente, el resultado del ordenamiento mostrará al archivo ordenado según sexo y dentro de sexo los valores de edad se dispondrán en forma descendente.

Los botones, en la parte inferior de la ventana de diálogo, permiten cambiar el criterio de orden (ascendente o descendente) y la jerarquía del ordenamiento.

Por ejemplo, utilizando los datos del archivo *Atriplex*, se eligió ordenar las observaciones en orden descendente, de acuerdo a los valores de la variable PG. La configuración resultante se muestra en la siguiente tabla:

Tabla 1: Archivo Atriplex ordenado en orden descendente según la variable PG.

Tamaño	Episperma	PG	PN	PS
medianas	rojizo	100	80	0.0032
grandes	claro	93	80	0.0040
medianas	claro	93	80	0.0038
medianas	claro	93	80	0.0043
chicas	rojizo	93	7	0.0030
grandes	claro	87	87	0.0043
medianas	claro	87	54	0.0033
.	.	.	.	.
.	.	.	.	.
chicas	oscuro	20	0	0.0030
medianas	oscuro	13	7	0.0030

Opcionalmente la acción de ordenar se puede invocar desde la barra de herramientas activando el ícono **Ordenar**.

Advertencia: esta opción no puede deshacerse automáticamente. Para mantener el archivo original, cerrar la tabla sin guardar este cambio, guardar el archivo con otro nombre o bien realizar el ordenamiento que recupere la disposición original de los datos.

Categorizar

Menú DATOS ⇒ CATEGORIZAR, permite categorizar los datos de una **columna previamente seleccionada**, generando una nueva columna con la categorización. Esta acción estará disponible si en la columna seleccionada los datos son de naturaleza *entera* o *real*. Hay dos procedimientos disponibles: **asignar categorías por intervalos** o **asignar categorías por una tabla de asignación**.

Al elegir **asignar categorías por intervalos**, la categorización se obtiene estableciendo los límites superiores de un conjunto de intervalos de clase. Los casos que pertenecen a cada clase reciben la misma categoría. Según la forma en que se establecen los límites de los intervalos de clase, se definen los siguientes métodos de categorización:

FIJO: categoriza el conjunto de datos generando tantos intervalos como el número de categorías solicitado. Se muestran el valor mínimo, el valor máximo, la longitud del intervalo y los límites superiores de cada categoría identificados como C1, C2, etc. Si se desea representar las categorías con números enteros se debe activar la casilla **Numéricas**. Por defecto el orden de las categorías es ascendente, para cambiarlo se debe activar la casilla **Descendente**. Para ejecutar la categorización presione el botón **Aceptar**. Los valores **Mínimo** y **Máximo** pueden cambiarse para lograr una categorización de acuerdo al requerimiento del usuario.

PROBABILÍSTICO: el límite superior de cada categoría representa un percentil de la distribución de la variable de acuerdo al número de intervalos solicitados. Por ejemplo, si se solicitan 4 intervalos los límites para ellos serán respectivamente, los percentiles 25, 50, 75 y 100. Para aplicar la categorización presione el botón **Aceptar**.

PERSONALIZADO: se pueden ingresar los límites superiores de los intervalos de cada categoría. Para ello elija un número de categorías que se desee crear y en la tabla adyacente ingrese los límites superiores de cada una de ellas. Por defecto la última categoría presenta como límite superior el máximo de los valores observados. Para aplicar la categorización presione el botón **Aceptar**.

A modo de ejemplo, utilizando los datos del archivo *Atriplex* (previamente ordenado según un criterio descendente para la variable PG), se eligió categorizar las observaciones asignando categorías por intervalos. La configuración resultante se muestra en la Tabla 2. Usando la opción FIJO, se seleccionó la configuración preestablecida: Nº Categorías: 5, mín: 13; máx: 100; Longitud del intervalo: 17.4; límites superiores de los intervalos: 30.4; 47.8; 65.2; 82.6; 100. Usando la opción PROBABILÍSTICO también se seleccionaron 5 categorías y los límites superiores de los intervalos fueron: 33; 60; 80; 87 y 100. Usando la opción PERSONALIZADO, se eligieron dos categorías: una con valores de germinación menores e iguales a 80% para lo cual en el campo **LS1** hay que escribir 80, y otra con valores mayores a 80%, lo que ya está contemplado en el campo **LS2** que contiene el valor 100 por defecto.

Tabla 2: Archivo Atriplex con la variable PG categorizada según tres criterios.

Germ.	Fijo	Prob.	Pers.		Germ.	Fijo	Prob.	Pers.
100.00	C5	C5	C2		73.00	C4	C3	C1
93.00	C5	C5	C2		66.00	C4	C3	C1
93.00	C5	C5	C2		60.00	C3	C2	C1
93.00	C5	C5	C2		60.00	C3	C2	C1
93.00	C5	C5	C2		53.00	C3	C2	C1
87.00	C5	C4	C2		53.00	C3	C2	C1
87.00	C5	C4	C2		40.00	C2	C2	C1
87.00	C5	C4	C2		33.00	C2	C1	C1
87.00	C5	C4	C2		33.00	C2	C1	C1
87.00	C5	C4	C2		26.00	C1	C1	C1
80.00	C4	C3	C1		20.00	C1	C1	C1
80.00	C4	C3	C1		20.00	C1	C1	C1
80.00	C4	C3	C1		13.00	C1	C1	C1
73.00	C4	C3	C1					

Al elegir **asignar categorías por una tabla de asignación**, las categorías se pueden leer desde una tabla o pueden ser ingresadas por el usuario. Este procedimiento es útil, por ejemplo, en el caso de tener un archivo en el que se utilizó codificación numérica para representar los diferentes estados de variables de naturaleza cualitativa. La ventana de diálogo se muestra a continuación.

En la tabla de diálogo, a la izquierda se muestra la lista de números encontrados en la variable a categorizar y, a la derecha, una lista vacía de categorías. Las categorías se ingresan manualmente, o se leen desde una tabla contenida en un archivo de texto o desde una tabla residente en la memoria. El archivo de texto debe contener tantos renglones como categorías y en cada renglón debe figurar un número seguido de un signo separador (puede ser "=", ":", ";" o un tabulador), y a continuación el nombre de la categoría asociada a ese número. Por ejemplo, si al registrar el tipo de ocupación el número 2 corresponde a la categoría desempleado, esto deberá figurar como: 2=desempleado). Si se utiliza la opción de asignar las categorías por una tabla almacenada en la memoria (*clipboard*), esta tabla debe haber sido copiada desde un archivo con la estructura descrita para los archivos de texto. Estas opciones de carga se seleccionan desde un menú que aparece apretando el botón derecho del ratón sobre la tabla de asignaciones, como se muestra en la figura. Para que esté disponible la opción **Copiar desde la memoria**, la tabla debe estar en memoria.

Para obtener la categorización debe presionar el botón Aceptar. Las categorías estarán contenidas en una nueva columna con una etiqueta que lleva el prefijo Cat y luego el nombre de la variable a la cual corresponde la categorización. En la figura se puede ver un campo de edición en el que se lee Cat_Ocupación, lo que puede ser modificado escribiendo un nuevo nombre.

Cuando una variable numérica es categorizada usando una tabla de asignación, en la descripción de la variable resultante se puede leer dicha tabla.

Editar categorías

Para aplicar esta acción se debe tener seleccionada la columna que contiene las categorías. Menú DATOS ⇒ EDITAR CATEGORÍAS, presenta una ventana de diálogo (**Editar categorías**) que muestra las categorías de la variable (columna) seleccionada. En esta ventana aparecerá una lista que contiene las categorías actuales. Seleccionando una categoría, su nombre aparece en un campo de edición que está por encima de la lista. En ese campo se puede modificar el nombre de la categoría. Este cambio se visualiza en la lista

automáticamente. Si se acciona el botón **Aceptar**, los cambios se reflejarán en la tabla de datos.

Una categoría puede ser amalgamada con otra ubicándola con los botones con flechas: **Nivel superior** y **Nivel inferior**. Si se selecciona una categoría y se presionan la flecha de dirección derecha (**Nivel inferior**) la categoría seleccionada quedará "incluida" dentro de la categoría que le sigue en la lista. Si se acciona el botón **Aceptar** la categoría incluida desaparecerá de la tabla de datos siendo sustituida por la categoría que la incluye. Otra forma de incluir una categoría dentro de otra es seleccionarla con el ratón, y con el botón del ratón apretado, arrastrarla hasta la posición donde se encuentra la categoría que la va a contener. Si una categoría es mal incluida marcándola con el ratón se la puede arrastrar hasta la nueva categoría en donde se desea incluirla. Antes de presionar el botón **Aceptar** la acción de inclusión puede ser revertida seleccionando la categoría incluida y presionando el botón con la flecha de dirección izquierda (**Nivel superior**). Para cambiar de posición las categorías se pueden usar las flechas **Subir orden** y **Bajar orden**. Una vez que se esta de acuerdo con la recategorización realizada, presionar el botón **Aceptar** para reflejar los cambios en la planilla de datos.

A los fines de facilitar el ingreso de datos de variables *categóricas* cada categoría tiene asociado un número de orden que depende de su posición en la lista que aparece en la ventana de diálogo **Editar categorías**. Por ejemplo si las categorías son "pequeña", "mediana" y "grande" y aparecen en la lista en ese orden, con sólo ingresar un 1 en una celda de la columna que contiene a estas categorías, al dar <Enter> aparecerá el nombre "pequeña". Si se altera el orden de las categorías en la lista la codificación numérica responderá al nuevo ordenamiento.

Si se cambia el tipo de una variable de *categórico* a *entero,* se obtendrán números que se corresponden con el orden de la categoría en la lista.

En la barra de herramientas de InfoStat, se encontrará el botón que figura en este párrafo, el cual permite editar categorías sin necesidad de ir al Menú DATOS. 

Transformar

Al invocar esta acción aparece la ventana **Transformaciones** para elegir la o las variables a transformar; éstas deben ser variables cuantitativas. Luego de **Aceptar**, aparece otra ventana que permite elegir la transformación. En esta ventana aparecen dos listas de transformaciones: una para aplicar a una variable y otra para aplicar a combinaciones de variables. Cualquiera sea la transformación elegida InfoStat generará nuevas columnas conteniendo las variables transformadas, a las que nombrará automáticamente con el nombre de la transformación y después de un subguión el nombre de la variable original.

Selección de la transformación: las transformaciones posibles son: **Estandarizar**, **Estandarizar por filas**, **Centrar**, **Centrar por fila**, **Residuos Ext Estud**. (residuos externamente estudentizados), **Rangos**, **Escores normales**, **Log10** (logaritmo en base 10), **Log2** (logaritmo en base 2), **Ln** (logaritmo natural), **Raíz cuadrada**, **Recíproco**, **Potencia**, **ArcoSeno(Raíz(p))** (arcoseno de la raíz cuadrada de *p*, donde *p*=valor de la variable expresado en proporción), **Probit**, **Logit**, **Complemento log-log, Llevar al intervalo (0,1),**

1 si >=media sino 0, 1 si >=mediana sino 0 y **Acumular**. Si se eligen dos o más variables se pueden obtener otras transformaciones que figuran en la lista **Combinación de variables**.

Estandarizar: permite obtener la estandarización de la o las variables seleccionadas. La estandarización se realiza sustrayendo de cada observación la media de la columna y dividiendo el resultado por la desviación estándar de los valores en la columna.

Estandarizar por filas: si el usuario selecciona más de una variable en el menú transformar, se habilita la opción estandarizar por filas. En este caso cada entrada en la tabla es transformada a su valor estandarizado con la media y desviación estándar de los elementos de la fila correspondiente.

Centrar: esta transformación realiza un centrado por columna. Es decir, a cada observación de la variable seleccionada, se le sustrae el valor de la media de dicha variable obtenida con los datos de la correspondiente columna.

Centrar por filas: en este caso a cada valor de una variable seleccionada se le sustrae la media obtenida por fila con los datos de todas las variables que fueron seleccionadas.

Escores normales: a la variable seleccionada se le aplica la transformación rango. Luego, cada valor de rango es dividido por (n+1), siendo *n* el total de datos de la muestra. Para cada cociente se obtiene la función de distribución inversa correspondiente a una Normal (0;1).

Residuos Ext Estud. (residuos externamente estudentizados): para un modelo de posición se definen como:

$$REE = \left(y_i - \overline{y}^{(-i)} \right) \Big/ S^{(-i)}$$

donde y_i es el valor de la observación que no se considera, $\overline{y}^{(-i)}$ es la media de los datos sin la observación y_i, siendo $S^{(-i)}$ es el desvío estándar de los datos calculado después de la eliminación de la observación.

Rangos: esta función asigna a los datos originales la posición que cada uno ocupa en la serie ordenada en forma ascendente. En un grupo de *n* datos al menor le corresponde el rango 1, al segundo más pequeño el rango 2 y así sucesivamente. El valor más alto tendrá el rango *n*. Si dos o más observaciones muestran un mismo valor (empate), el rango asignado a cada una es el promedio de los rangos consecutivos correspondientes a ese valor.

Por ejemplo para la serie 10, 20, 20, 30, 40, 50, 50, 50, 60; la serie transformada es: 1, 2.5, 2.5, 4, 5, 7, 7, 7, 9.

Transformación logaritmo: InfoStat permite generar variables a partir de las funciones **Log10** (logaritmo en base 10), **Log2** (logaritmo en base 2) y **Ln** (logaritmo natural). Si el valor a transformar es menor o igual que cero el resultado es un valor faltante. En este caso se puede usar *log(y+c),* donde *c* es una constante.

Raíz cuadrada: $\sqrt{y}$ o bien $\sqrt{y+c}$ donde *c* es una constante.

Recíproca: $1/y$.

Potencia: y^{λ} con $\lambda \neq 0$ donde λ es la potencia deseada.

ArcoSeno(Raíz(p)): $Sen^{-1}\left(\sqrt{p}\right)$ con $p \in [0,1]$ (arcoseno de la raíz cuadrada de la proporción)

Probit: se define como $Probit\ (p)=\Phi^{-1}(p)$ con $p \in (0,1)$, donde Φ^{-1} es la inversa de la función normal acumulada.

Logit: se define como $Logit\ (p)=ln(p/(1-p))$ con $p \in (0,1)$.

Complemento log-log: se define como $CLL(p)=ln[-ln(1-p)]$ con $p \in (0,1)$.

Llevar al intervalo (0,1): dado un conjunto $\{y_1,...,y_n\}$ de observaciones, la transformación consiste en restarle a cada valor el mínimo de $\{y_1,...,y_n\}$ y dividirlo por el recorrido (diferencia entre el máximo y el mínimo).

1 si >=media sino 0: permite dicotomizar los datos en función de la media de las observaciones. Los datos mayores o iguales que la media tendrán valor 1.

1 si >=mediana sino 0: permite dicotomizar los datos en función de la mediana de las observaciones. Los datos mayores o iguales que la mediana tendrán valor 1.

Acumular: Genera una columna donde el elemento t-ésimo representa la suma de los primeros *t* elementos. Por ejemplo si la columna contiene los valores 10, 12 y 20, aplicando esta opción se obtendrá 10, 22 y 42.

Combinación de variables permite aplicar funciones que involucran varias columnas del archivo. En el selector de variables se deberán especificar las variables que intervienen en la evaluación de la función seleccionada. La función a seleccionar puede ser una de las siguientes: **Suma**, **Media**, **Mediana**, **Varianza**, **Desviación estándar**, **Mínimo**, **Máximo** y **Combinación lineal.** La función **Suma** realizará la suma de los valores de las columnas seleccionadas en cada fila del archivo y generará una nueva variable que se denomina *Suma*. De igual manera, se puede solicitar la Media, Mediana, Varianza, Desviación Estándar, Mínimo y Máximo de los valores en cada fila. Cuando se selecciona combinación lineal se deben indicar los coeficientes de la combinación en la ventana **Coeficientes**. Los coeficientes se deben ingresar de a uno por vez dando <Enter>. Así, si se tienen dos columnas, digamos X e Y, y se especifican los números 2 y 3 en la ventana coeficientes, se generará una nueva columna denominada combinación lineal igual a 2X+3Y.

Crear variables auxiliares (dummy)

En algunas aplicaciones estadísticas, por ejemplo aquellas relacionadas a modelos de regresión, es necesario transformar una variable categórica *X* con *k* categorías en *k-1* variables binarias (con valor 0 ó 1). Una variable binaria de este tipo es conocida con el nombre de variable auxiliar o variable dummy. El conjunto de *k-1* variables auxiliares es utilizado para identificar cada una de las categorías de la variable original *X*. Así por ejemplo, si *X* tiene *k*=3 categorías, dos variables auxiliares D1 y D2 serán suficientes para representar cada una de las categorías de *X*. Por ejemplo, la combinación D1=1 y D2=0 puede identificar la primera categoría, D1=0 y D2=1 la segunda categoría y D1=0 y D2=0 la

tercera categoría. En este caso, a la tercera categoría (aquella donde todas las variables auxiliares asumen el valor cero) se la suele llamar categoría de referencia.

Para crear variables auxiliares, seleccionar la variable categórica original, al **Aceptar**, aparecerá la pantalla **Generador de variables auxiliares** donde se listará la o las variables originales y las categorías disponibles para cada una de ellas. La primera categoría aparecerá automáticamente seleccionada para ser usada como categoría de referencia. Si el usuario desea que otra sea la categoría de referencia deberá mover el curso hasta esa categoría para seleccionarla. InfoStat generará las *k-1* variables auxiliares, que se agregarán a la tabla de datos, a las que denominará con el nombre de la variable original seguidos por una extensión para su diferenciación.

La opción **Multiplicar por…** que aparece en la pantalla **Generador de variables auxiliares** sirve para obtener el producto entre las variables auxiliares y alguna variable de interés. Dichos productos se mostrarán en nuevas columnas de la tabla de datos, con un nombre que indique su origen. Un ejemplo de aplicación de esta opción puede consultarse en **Regresión con variables auxiliares**.

Llenar con...

El llenado automático completa un conjunto de celdas seleccionadas según la opción de llenado especificada. Para llenar celdas, selecciónelas y del menú principal elija DATOS ⇒ LLENAR CON... y especifique el tipo de llenado.

Advertencia: estas acciones reemplazan los valores de la columna seleccionada, por lo que si se quiere preservar el contenido de la columna original se deberá duplicar la misma y aplicar la distribución sobre ésta.

Completando hacia abajo

Las celdas vacías reciben el contenido de la primera celda no vacía que las antecede en la misma columna. Esta acción también se puede realizar con las teclas CTRL+D.

Con secuencia 1, 2,...

Las celdas seleccionadas, comenzando desde la primera celda seleccionada, reciben un número natural con una secuencia en sentido ascendente y siguiendo con las columnas de la derecha sin volver la numeración al punto inicial cuando se cambia de columna.

Con Uniforme (0,1)

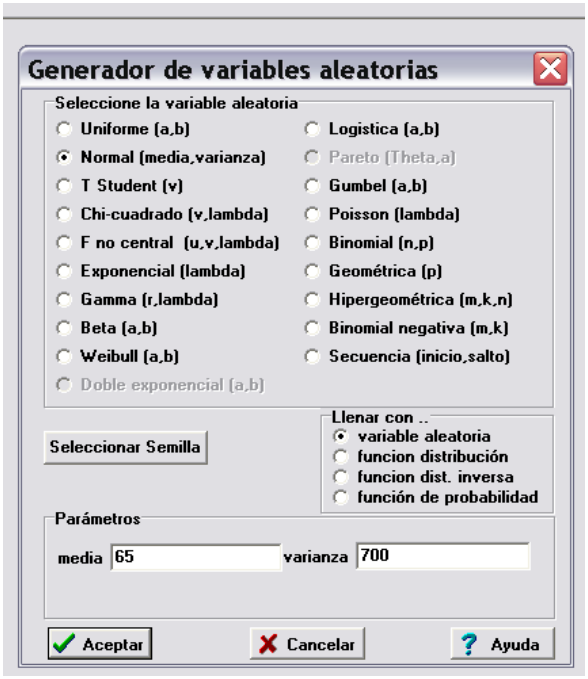
Al elegir esta opción las celdas seleccionadas recibirán un valor de una variable aleatoria continua con distribución uniforme, entre 0 y 1.

Con Normal (0,1)

Al elegir esta opción las celdas seleccionadas serán reemplazadas con realizaciones de una variable aleatoria con distribución normal con media = 0 y varianza = 1.

Otros...

Para una amplia lista de distribuciones de variables aleatorias, InfoStat permite llenar las celdas seleccionadas con: 1) realizaciones de **la variable aleatoria**, 2) **función de distribución** acumulada para argumentos leídos desde las celdas seleccionadas, 3) **función de distribución inversa**, evaluada de acuerdo a los valores seleccionados y 4) **función de probabilidad**, evaluada de acuerdo a los valores seleccionados.



Las distribuciones disponibles son: Uniforme, Normal, T de Student, Chi cuadrado, F no central, Exponencial, Gamma, Beta, Weibull, Logística, Gumbel, Poisson, Binomial, Geométrica, Hipergeométrica y Binomial negativa.

También se encuentra la opción **Secuencia (inicio, salto)**, con la que se podrá llenar celdas con una secuencia de números reales con un inicio y distancia entre dos números consecutivos a definir por el usuario en la subventana **Parámetros** (inicio y salto) que se habilita al seleccionar Secuencia (inicio-salto). Por ejemplo si el número de inicio es 1 y el salto de 2, la columna seleccionada comenzará con 1 seguirá con 3, luego con 5 y así sucesivamente.

Para llenar con realizaciones, función de distribución, función de distribución inversa o función de probabilidad de una de las variables aleatorias disponibles, seleccione la variable aleatoria y en el panel denominado **Parámetros**, especifique las constantes que caracterizan a la distribución elegida.

Seleccionar semilla: por defecto InfoStat utiliza una semilla aleatoria para generar números aleatorios, sin embargo en algunos casos es útil poder generar una misma secuencia aleatoria. Esto se logra especificando un mismo número arbitrariamente elegido, distinto de cero, en el campo de edición que se activa cuando se presiona el botón **Seleccionar semilla**. Si se pone como semilla el número cero, esto indica a InfoStat que la semilla es de origen aleatorio y por lo tanto las secuencias serán siempre diferentes.

A continuación se presenta una breve descripción de las distribuciones disponibles:

Nota: se designará como *E(X)* y *V(X)* a la esperanza y la varianza de la variable aleatoria (X) respectivamente.

Uniforme (a,b): Se dice que una variable aleatoria continua *X* tiene función de densidad uniforme en el intervalo [a,b] si:

$$f\ (x;a,b) = \frac{1}{b-a}\ \mathrm{I}_{[a,b]}(x)$$

con $\mathrm{I}_{[a,b]}(x)$ función indicadora, donde los parámetros *a* y *b* satisfacen $-\infty<a<b<\infty$. La $E(X)=(a+b)/2$ y $Var(X)=(b-a)^2/12$.

Normal (media,varianza): Una variable aleatoria continua *X*, con $-\infty< x<\infty$, está normalmente distribuida si su función de densidad viene dada por:

$$f\ (x;m,v) = \frac{1}{\sqrt{2\pi v}}\, e^{-(x-m)2\,/\,2v}$$

donde los parámetros *m* (media) y *v* (varianza) satisfacen $-\infty< m<\infty$ y $v>0$. InfoStat usa *m* y *v* para representar los parámetros $E(X)=\mu$ y $Var(X)=\sigma^2$ respectivamente.

T de Student (v): La variable aleatoria continua *X* (con $-\infty<x<\infty$) tiene una distribución conocida como T de Student con ν grados de libertad, si su densidad es:

$$f\ (x;\upsilon) = \frac{\Gamma\left[\left(\nu+1\right)/2\right]}{\Gamma(\nu/2)}\ \frac{1}{\sqrt{\nu\pi}}\ \frac{1}{\left(1+x^2/\nu\right)^{(\nu+1)/2}}$$

donde ν es un entero positivo conocido como grados de libertad y $\Gamma(.)$ es la función gamma, que tiene la siguiente forma:

$$\Gamma(r) = \int_0^{\infty} y^{r-1}\ e^{-y}\ dy$$

La $E(X)=0$ para grados de libertad mayor que 1 y $V(X)=\nu/(\nu-2)$ para $\nu>2$.

Chi cuadrado (v,lambda) (no central): La variable aleatoria *X* tiene distribución Chi cuadrado si su función de densidad es:

$$f(x;\nu,\lambda) = \sum_{j=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^{j}}{j!}\right) \left(\frac{x^{(\nu+2j-2)/2}\ \ e^{-x/2}}{\Gamma\left(\frac{\nu+2j}{2}\right) 2^{j+(\nu/2)}}\right) \mathrm{I}_{(0,\infty)}(x)$$

con $\mathrm{I}_{(0,\infty)}(x)$ función indicadora, ν un entero positivo conocido como grados de libertad, $\Gamma(.)$ es la función gamma, $\lambda\geq 0$ conocido como parámetro de no centralidad, y definiendo $\lambda^j=1$ cuando $\lambda=0$, $j=0$.

La $E(X)=\nu+2\lambda$ y la $V(X)=2(\nu+4\lambda)$. Si $\lambda=0$ la distribución es conocida como **Chi cuadrado central.**

F no central (u,v,lambda): La variable aleatoria continua *X* tiene distribución *F no central* caracterizada por sus grados de libertad *u* (grados de libertad para el numerador) y ν (grados de libertad para el denominador) y por el parámetro de no centralidad, λ , si su función de densidad está dada por:

$$f(x;u,\nu,\lambda)=\sum_{j=0}^{\infty}\frac{\lambda^j e^{-\lambda}\ \Gamma\left(\frac{2j+u+\nu}{2}\right)\left(\frac{u}{\nu}\right)^{(u+2j)/2}x^{(u+2j-2)/2}}{j!\ \Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{2j+u}{2}\right)\left(1+\frac{ux}{\nu}\right)^{(u+\nu+2j)/2}}\ \mathrm{I}_{(0,\infty)}(x)$$

con $\mathrm{I}_{(0,\infty)}(x)$ función indicadora, *u* y ν enteros positivos, *Γ(.)* es la función, $\lambda \geq 0$ definiendo $\lambda^j=1$ cuando $\lambda=0$, $j=0$. Si $\lambda=0$ la distribución es conocida como F central y su $E(X)=\nu/\nu-2$ para $\nu>2$ y la $V(X)=2\ \nu^2(u+\nu-2)/u(\nu-2)^2(\nu-4)$ para $\nu>4$.

Exponencial (lambda): La variable aleatoria continua *X* tiene distribución exponencial si su función de densidad está dada por:

$$f\ (x;\lambda)=\lambda e^{-\lambda x}\ \mathrm{I}_{(0,\infty)}(x)$$

con $\mathrm{I}_{(0,\infty)}(x)$ función indicadora y $\lambda>0$. La $E(X)=1/\lambda$ y $V(X)=1/\lambda^2$.

Gamma (r,lambda): La variable aleatoria continua *X* tiene distribución gamma, si su función de densidad está dada por:

$$f\ (x;r,\lambda)=\frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x}\ \mathrm{I}_{(0,\infty)}(x)$$

con $\mathrm{I}_{(0,\infty)}(x)$ función indicadora, $r>0$ y $\lambda>0$ y donde *Γ(.)* es la función gamma. La $E(X)=r/\lambda$ y $V(X)=r/\lambda^2$.

Beta (a,b): La variable aleatoria continua *X* tiene distribución beta si su función de densidad está dada por:

$$f\ (x;a,b)=\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}\ \ \mathrm{I}_{(0,1)}(x)$$

con $\mathrm{I}_{(0,1)}(x)$ función indicadora, $a>0$, $b>0$ y $B(a,b)$ es la función beta, dada por la siguiente expresión:

$$B(a,b)=\int_0^2 x^{a-1}(1-x)^{b-1}\ dx \quad para\ a>0,\ b>0$$

La $E(X)=a/(a+b)$ y $V(X)=ab/((a+b+1)(a+b)^2)$.

Weibull (a,b): La variable aleatoria *X* tiene distribución Weibull si su función de densidad es:

$$f(x;a,b) = abx^{b-1}e^{-ax^b}\ \mathrm{I}_{(0,x)}(x)$$

con $\mathrm{I}_{(0,x)}(x)$ función indicadora, a>0 y b>0. La $E(X)=(1/a)^{1/b}\ \Gamma(1+b^{-1})$ y $V(X)=(1/a)^{2/b}\ [\Gamma(1+2b^{-1})-\Gamma^2(1+b^{-1})]$, donde $\Gamma(.)$ es la función gamma.

Logística (a,b): La variable aleatoria *X* tiene distribución logística si su función de distribución acumulada está dada por:

$$F\left(x;a,b\right) = \left[1+e^{-(x-a)/b}\right]^{-1}$$

donde $-\infty<a<\infty$ y $b>0$. La $E(X)=a$ y la $V(X)=(\pi^2b^2)/3$.

Gumbel o valor extremo (a,b): La variable aleatoria *X* tiene distribución Gumbel si su función de distribución acumulada está dada por:

$$F(x;a,b) = exp(-e^{-(x-a)/b})$$

donde $-\infty<a<\infty$ y $b>0$. La $E(X)=a-b\gamma$ donde γ se aproxima a 0.577216 y $V(X)=(\pi^2b^2)/6$.

Poisson (lambda): Esta distribución da un modelo para variables de tipo conteo, donde los conteos se refieren al registro del número de eventos de interés en una unidad de tiempo o espacio dados (horas, minutos, m^2, m^3, etc.). Se dice que una variable aleatoria discreta *X* tiene distribución Poisson si su función de densidad está dada por:

$$f\left(x;\lambda\right) = \frac{e^{-x}\lambda^x}{x!}\ \mathrm{I}_{[0,1,\ldots]}(x)$$

con $\mathrm{I}_{[0,1,\ldots]}(x)$ función indicadora y el parámetro $\lambda>0$. La $E(X)=\lambda$ y $Var(X)=\lambda$.

Binomial (n,p): Esta distribución tiene origen cuando ocurren las siguientes condiciones en forma simultánea: a) se realizan *n* ensayos Bernoulli, b) el parámetro *p* (probabilidad de "éxito") se mantiene constante entre ensayos y c) los ensayos son independientes entre sí.

Distribución Bernoulli: en ciertos experimentos puede ocurrir que existan sólo dos resultados posibles: éxito o fracaso, presencia o ausencia, si o no, etc. Una variable Bernoulli es aquella variable binaria que identifica estos eventos. Por ejemplo, se puede tomar $x=1$ para representar el éxito y $x=0$ para representar al fracaso. La $E(X)=p$ y la $V(X)=p(1-p)$, donde *p* es la probabilidad de éxito.

Se dice que una variable aleatoria discreta *X* tiene distribución Binomial si su función de densidad está dada por:

$$f(x;n,p) = \binom{n}{x} p^x q^{n-x}\ \mathrm{I}_{[0,1,\ldots,n]}(x)$$

con $\mathrm{I}_{[0,1,\ldots,n]}(x)$ función indicadora y el parámetro $0\leq p\leq 1$, $q=1-p$ y $n=1,2,\ldots$ es el total de ensayos. La $E(X)=np$ y $Var(X)=npq$.

Geométrica (p): Esta distribución es de especial interés en la modelización del *número de ensayos necesarios hasta que aparezca el primer éxito.* Una variable aleatoria discreta *X* tiene distribución geométrica (o de *Pascal*) si su función de densidad está dada por:

$$f\left(x;p\right)=p\left(1-p\right)^{x}\ \mathrm{I}_{[0,1,...]}\left(x\right)$$

donde $\mathrm{I}_{[0,1,...]}\left(x\right)$ función indicadora y el parámetro $0\leq p\leq 1$, $q=1-p$. La $E(X)=q/p$ y $Var(X)=q/p^{2}$.

Hipergeométrica (m,k,n): Esta distribución está ligada a situaciones de muestreo sin reposición, es decir, situaciones en que se elige al azar un elemento de una población y así sucesivamente hasta completar la muestra sin restituir los elementos extraídos. Considérese como población a un conjunto de *m* elementos de los cuales *k* poseen uno de dos estados posibles (éxito) y *m-k* presentan el otro (fracaso). Al igual que en la distribución Binomial el problema de interés es hallar la probabilidad de obtener *x* éxitos en una muestra de tamaño *n*. Una variable aleatoria discreta *X* tiene distribución Hipergeométrica si su función de densidad está dada por:

$$f\ \left(x;m,k,n\right)=\frac{\binom{k}{x}\binom{m-k}{n-x}}{\binom{m}{n}}\ \mathrm{I}_{[0,1..,n]}\ \left(x\right)$$

donde $\mathrm{I}_{[0,1,...]}\left(x\right)$ es una función indicadora, el parámetro $m=1,2,...,$ el parámetro $k=0,1,...m$ y $n=1,2,...,m.$ La $E(X)=n(k/m)$ y $Var(X)=n(k/m)\ ((m-k)/m)\ ((m-n)/m-1)$.

Binomial negativa (m,k): En conexión con la repetición de ensayos Bernoulli, ciertos problemas, comunes en estudios de poblaciones naturales, centran su atención en la probabilidad de encontrar *x* individuos en una unidad muestral bajo situaciones donde los individuos tienden a estar agregados (distribución de contagio). InfoStat permite calcular esas probabilidades a través de la función Binomial negativa. Se dice que una variable aleatoria discreta *X* tiene distribución Binomial negativa si su función de densidad está dada por:

$$f\left(x;m,k\right)=\left(\frac{1}{q^{k}}\right)\left(\frac{\left(k\right)\left(k+1\right)\left(k+2\right)...\left(k+x-1\right)}{x!}\right)\left(\frac{p}{q}\right)^{x}\ \mathrm{I}_{[0,1,...]}\left(x\right)$$

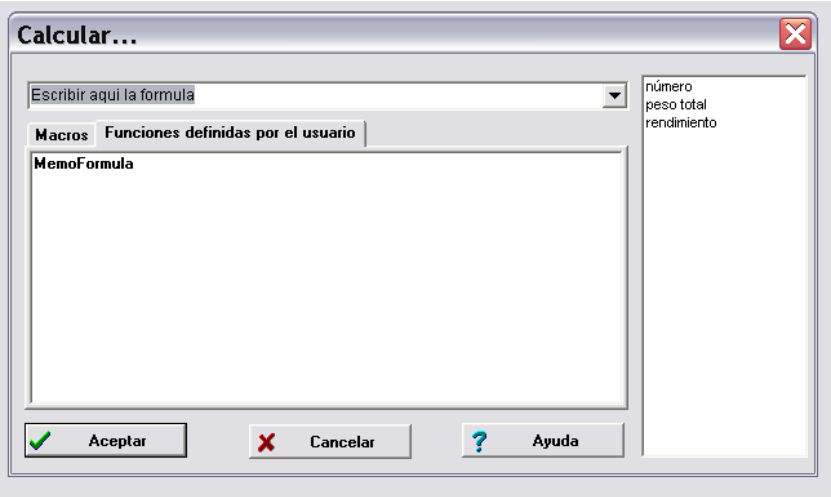
donde $\mathrm{I}_{[0,1,...]}\left(x\right)$ es una función indicadora, $p=m/k$ y $q=p+1$. Los parámetros *m* y *k* satisfacen: $m>0$ (número promedio de individuos por unidad muestral) y $k>0$ (parámetro de contagio o agregación).

Fórmulas

Permite especificar una fórmula cuyos resultados pueden sustituir el contenido de una columna existente o ser contenidos en una nueva.

Advertencia: los nombres de las variables involucradas en el cálculo no deben tener paréntesis, símbolos de operadores matemáticos o nombre de funciones reservadas, pero si pueden contener acentos y eñes.

La ventana de diálogo se muestra a continuación:



Durante una sesión de trabajo, las fórmulas que se van escribiendo quedan almacenadas en una lista y están disponibles para volver a utilizarlas. Para visualizarlas haga *click* sobre la esquina derecha del campo donde escribe las fórmulas.

La ventana de diálogo muestra una lista de las variables disponibles, las cuales pueden ser incluidas en la fórmula haciendo clic sobre el nombre en la lista. Cuando se utiliza este procedimiento para agregar las variables a la expresión que se está escribiendo, los nombres aparecen enmarcados entre corchetes. Esto permite incluir en una fórmula nombres que contienen espacios o símbolos matemáticos que no deben ser interpretados como tales.

Se pueden usar funciones predefinidas o el usuario puede definir sus propias funciones. Para el último caso debe escribir la función en el panel que aparece por debajo del campo de edición de fórmulas. Por ejemplo, la función cubo(x) no es una función predefinida pero puede ser especificada por el usuario en el panel **Funciones definidas por el usuario** escribiendo: cubo(x)=x*x*x. Esta definición permitirá aplicar la función cubo a cualquier otra variable de la tabla activa o a cualquier otra expresión válida. Escribiendo en el campo de especificación de las fórmulas por ejemplo h=cubo(COLUMNA1), se aplicará la función cubo a los datos de la columna 1.

Si las variables involucradas en la fórmula tienen nombre muy largo se pueden sustituir esos nombres, en la fórmula, con %# donde # es el número de la columna donde se encuentra la variable. Por ejemplo, si la tabla de datos posee 3 columnas, %1 representará al nombre de la primer columna, %2 al de la segunda columna y %3 al nombre de la tercera. Para identificar las correspondencias entre el nombre y el número de columna se debe presionar la tecla **Alt**. Mientras esta tecla se mantenga apretada los nombres de las columnas de la tabla activa serán mostrados como %#.

Si se desea aplicar una función como media(.), min(.), max(.), que aceptan múltiples argumentos, a un bloque de variables debe usarse la notación f(%a**:**%b) donde f denota la función, %a y %b indican el número de la columna del comienzo y fin del bloque, respectivamente. Nótese que el carácter que separa el comienzo y fin de un bloque es "dos puntos" (:). Siguiendo con el ejemplo de arriba, para calcular el promedio de las 3 primeras variables del archivo se indicará: media (%1**:**%3). Otra forma de indicar que la función se aplicará a un conjunto de variables como, por ejemplo, **media** (), es usar el formato **media** (*nombre variable1:nombre variableN*) indicando que se quiere obtener la media de todas las variables entre la variable 1 y la n-ésima variable. Esta expresión se puede escribir manualmente o se escribe automáticamente si se selecciona, en la lista de variables, el bloque de variables.

Las tablas de datos IDB2 guardan las fórmulas que dan origen al contenido de una columna. Es posible actualizar el contenido de la columna aplicando nuevamente la fórmula. Para esto hay que seleccionar la columna y elegir la **opción Actualizar** del menú Datos o del menú que se despliega al presionar el botón derecho del ratón. Aparece el diálogo en el modo **Macros** con la correspondiente fórmula (o las fórmulas, si se seleccionó más de una columna). Estas fórmulas se pueden editar o ejecutar, selectiva o conjuntamente, para actualizar el contenido de la columna.

Se pueden efectuar modificaciones en la tabla de datos manteniendo abierta la ventana de fórmulas.

Para especificar una fórmula, elija en el menú DATOS ⇒ FÓRMULAS y en la ventana escriba la expresión, por ejemplo: Y=LN(COLUMNA1)+3.

Los operadores y funciones predefinidas en InfoStat son:

+ : operador suma.

- : operador resta.

*** :** operador multiplicación.

/ : operador división.

^ : operador exponente (solo números positivos en la base).

(: abrir paréntesis.

) : cerrar paréntesis.

e : constante 2.7172…

PI: constante 3.141592653…

ABS(x): valor absoluto de x (Rango de x: -1e4932...1e4932).

ARCOCOSENO(x) ó **ARCCOSIN**(x): Arcocoseno de x.

ARCOSENO (x) ó **ARCSIN** (x):: Arco seno de x.

AREAY(y1;…;yn): Calcula el área bajo la curva definida por los pares ordenados (Y,X) suponiendo que los valores de X están igualmente espaciados a una unidad.

AREAYX(y1;x1;…;yn;xn): Calcula el área bajo la curva definida por los pares ordenados (Y,X).

ATAN(x): Arco tangente de x (Rango de x: -1e4932...1e4932).

COSENO(x) ó **COS**(x): Coseno de x (Rango de x: -1e18...1e18).

CUADRADO(x) ó **SQR**(x): cuadrado de x (Rango de x: -1e2446... 1e2446).

DESVIO(x1;x2;…;xn) ó **STDEV**(x1;x2;…;xn: Calcula la desviación estándar de los valores de las variables indicadas.

DISTNORMAL(x;m;v): Calcula la probabilidad acumulada hasta x para una normal con media m y varianza v.

EXP(x): exponencial e^x (Rango de x: -11356...11356).

FACTORIAL(x): número factorial de x.

GAMMA(x): Asigna a los valores de la variable indicada, los valores de la función gamma.

INVNORMAL(p;m;v): Calcula el valor de x tal que la P(X<x)=p con X~N(m,v).

LN(x): logaritmo natural de x (Rango de x: 0...1e4932).

LN2(x): logaritmo en base 2 de x.

LOG10(x): logaritmo en base 10 de x.

MAX(x1;x2;…;xn): Calcula el valor máximo del grupo de datos indicado.

MEDIA(x1;x2;…;xn) ó **MEAN**(x1;x2;…;xn): Calcula la media de los valores de las variables indicadas.

MEDIANA(x1;x2;…;xn) ó **MEDIAN**(x1;x2;…;xn: Calcula la mediana de los valores de las variables indicadas.

MIN(x1;x2;…;xn): Calcula el valor mínimo del grupo de datos indicado.

MOD(x) : operador modulo (aplicable a enteros solamente).

NORMA(x1;x2;…;xn): Calcula la norma del vector **x**.

NORMAL(m, v): Genera realizaciones de una variable aleatoria normal con media m y varianza v.

REDONDEO(x) ó **ROUND**(x): redondeo de x (Rango de x: -1e9...1e9).

RAIZ(x) ó **SQRT**(x): raíz cuadrada de x (Rango de x: 0...1e4932).

SENO(x) ó **SIN**(x):seno de x (Rango de x -1e18...1e18).

SUMA(x1;x2;…;xn) ó **SUM**(x1;x2;…;xn): Suma los valores de las variables que se indiquen.

TANGENTE (x): Tangente de x.

TRUNCADO(x) ó **TRUNC**(x): toma el valor entero de x (Rango de x: -1e9... 1e9).

URN. Genera realizaciones de una variable aleatoria uniforme.

UNIFORME(a, b): Genera realizaciones de una variable aleatoria uniforme en el intervalo (a, b).

VARIANZA(x1;x2;…;xn) ó **VARIANCE**(x1;x2;…;xn: Calcula la varianza de los valores de las variables indicadas.

ZRN: Genera realizaciones de una variable aleatoria normal estándar.

Para trabajar con datos de tipo *fecha*, se dispone de las funciones descriptas a continuación (entre paréntesis están los argumentos que la función espera).

DIADELCICLO(fecha,dia,mes): con esta sentencia se puede generar una columna que en cada celda contendrá el día del ciclo (en escala de 1 a 365), según corresponda a la fecha leída, teniendo en cuenta que el ciclo comienza en el día y mes especificados en el argumento. Por ejemplo, si en el campo para escribir una fórmula se ingresa día=DIADELCICLO(fecha,1,9), se genera una columna con el nombre día que contiene números enteros entre 1 y 365 cada uno correspondiente a la fecha indicada en el argumento, siendo el día "1" del ciclo el día primero de septiembre. Así, según este ejemplo, si en la columna fecha el dato es 18/09/07, en la columna día se corresponderá con el entero 18; si en fecha el dato es 03/10/07, en la columna día se corresponderá con el entero 33.

FECHADELDIADELCICLO(diadelciclo,dia,mes,año): devuelve la fecha correspondiente al dia del ciclo especificado, según el día, mes y año correspondientes a la fecha de origen del ciclo. Si el argumento del año se omite, toma el año actual. Esta función es la inversa de la función DIADELCICLO.

DIAJULIANO(fecha): genera una columna conteniendo el día juliano que corresponda a cada dato leído en la columna fecha.

AÑO(fecha): genera una columna conteniendo el año que corresponda a cada dato leído en la columna fecha.

MES(fecha): genera una columna conteniendo el mes que corresponda a cada dato leído en la columna fecha.

DIA(fecha): genera una columna conteniendo el día del mes que correspondan a cada dato leído en la columna fecha.

FECHA(día, mes, año): genera una columna conteniendo la fecha correspondiente al día, mes y año especificados.

Buscar

Menú DATOS ⇒ BUSCAR presenta una ventana de diálogo que permite, dentro de una parte de la tabla que haya sido seleccionada previamente, la búsqueda de números, categorías o fechas, iguales, mayores, menores y/o diferentes al especificado por el usuario. Estos valores se pueden reemplazar por otro, en cuyo caso se activa la casilla **Reemplazar**, o bien se pueden excluir del análisis activando **Desactivar caso,** o se pueden colorear las celdas si se activa la casilla **Colorear**. La búsqueda puede especificarse para un contenido

completo (si se activa la casilla **Celda completa**), o para algunos elementos dentro de un texto. Al final de cada reemplazo o desactivación el buscador muestra un reporte de la cantidad de casos encontrados o desactivados.

Remuestreo

Menú DATOS ⇒ REMUESTREO permite obtener muestras desde un conjunto de datos usando la técnica de *bootstrap*, *jackknife*, muestreo *aleatorio con reposición* o muestreo *aleatorio sin reposición*, conforme lo especifique el usuario. La opción *bootstrap* realiza un muestreo aleatorio con reposición generando muestras de tamaño *n* igual al tamaño de la muestra original, mientras que la opción *aleatorio con reposición* permite al usuario generar muestras de un tamaño distinto a *n*. Se debe indicar cuál es la columna para la que se obtendrán las muestras y, si lo hubiere, un criterio de clasificación y/o una partición. Luego, hay que seleccionar la técnica de muestreo (en el panel **Tipo de remuestreo**) y los valores que reportará el muestreo (panel **Guardar**). Si elige *bootstrap* debe ingresar el número de muestras a extraer en el campo **Muestras Bootstrap**; si elige un muestreo *aleatorio con* o *sin reposición* deberá ingresar la cantidad de muestras a generar (en **Nro. muestras**) y el tamaño de las mismas (en **Tamaño muestral**). Se pueden obtener los valores de la variable que conforman cada una de las muestras solicitadas (opción **Muestras**) o bien una o varias medidas resumen de cada muestra (opciones **Media, Mediana, Máximo, Mínimo, Recorrido, Varianza,** Desvío estándar **(D.E.), Error estándar,** Coeficiente de variación **(C.V), Suma, Suma de Cuadrados,** Mediana de los desvíos absolutos respecto de la mediana **(MAD),** Percentiles **(P01, P05, P10, P20, P25, P50, P75, P80, P90, P95 y P99), Kurtosis y Asimetría**). Los resultados se muestran en una nueva tabla. En el caso de solicitar los valores de la variable la nueva tabla tendrá una columna para cada muestra. Si se piden una o más medidas resúmenes la nueva tabla contendrá cada muestra en una fila y cada medida en una columna.

Colorear selección

Menú DATOS ⇒ COLOREAR SELECCION permite colorear un conjunto de celdas previamente seleccionadas. Cuando se colorea una variable, dicha variable aparece con el color en el listado del **Selector de variables**. Esta característica es útil, por ejemplo, si se usan colores para distinguir grupos de variables.

Unir tablas

Menú DATOS ⇒ UNIR TABLAS permite unir a la tabla activa dos o más tablas, **Horizontalmente** o **Verticalmente.** La unión se hace de a una tabla por vez.

La unión **horizontal** agrega columnas a la tabla activa para incluir la nueva información y requiere que se elija uno o más criterios de unión. Establecidos dichos criterios se mostrará una ventana de diálogo para seleccionar la tabla a unir y las columnas que se unirán (agregarán) a la tabla activa. Dicha ventana contiene una lista de las tablas abiertas en pantalla en la cual habrá que indicar la tabla a unir. Si la tabla no está en la lista se debe presionar el botón **Otra tabla** y abrir la tabla correspondiente, desde el lugar en que se

encuentre, la cual será agregada a la lista. Al seleccionar una tabla de la lista se mostrarán los nombres de sus columnas (variables) con casilleros activados, indicando las variables que se unirán a la tabla activa. El usuario podrá desactivar aquellos que no desea que participen del proceso. En el caso que ambas tablas tengan igual nombre de variables, al agregar la nueva información InfoStat colocará un número al final del nombre de la columna incluída, para así poder distinguirlos. Si el usuario quiere reemplazar en la tabla activa el contenido de las columnas de igual nombre deberá activar la casilla **Sobrescribir**.

Al realizar el proceso de unir tablas horizontalmente se encontró que se adicionan las columnas solicitadas pero no incluye la información de la tabla original.

La unión **vertical** agrega a la tabla activa nuevas filas para incluir la información de las columnas coincidentes y crea nuevas columnas para aquellas variables que no sean coincidentes. El procedimiento es similar al indicado para el caso de la unión horizontal, salvo que aquí no se requiere de un criterio de unión.

Ubicar columnas una debajo de la otra

Menú DATOS ⇒ UNIR COLUMNAS VERTICALMENTE une el contenido de dos o más columnas en una única columna. En la ventana de diálogo se deben seleccionar las columnas a unir (opción **Columnas**) y la unión se realizará según el orden de dicha selección. Se puede indicar también la copia de los registros de alguna columna de interés (opción **Copiar…**). Hay una opción para realizar el procedimiento de unión solo con los casos activos. Al **Aceptar** se genera una nueva tabla que muestra el resultado de la unión.

Reubicar filas como columnas

Menú DATOS ⇒ REUBICAR FILAS COMO COLUMNAS permite trasladar el contenido de las filas de la tabla activa a columnas de una nueva tabla de acuerdo a un criterio de clasificación establecido por el usuario. En la ventana de diálogo en la opción **Columnas** se deberán indicar las variables cuyos datos se colocarán en las columnas de la nueva tabla y en la opción **Criterio de clasificación** aquellas variables que definirán las columnas de la nueva tabla. Se puede indicar también la copia de los registros de alguna columna de interés (opción **Copiar…**). Presionando **Aceptar** se obtendrá la tabla nueva.

Crear nueva tabla con los casos activos

Menú DATOS ⇒ CREAR NUEVA TABLA CON LOS CASOS ACTIVOS genera una nueva tabla que contendrá solamente los casos activos de una tabla en uso que contenga casos desactivados.

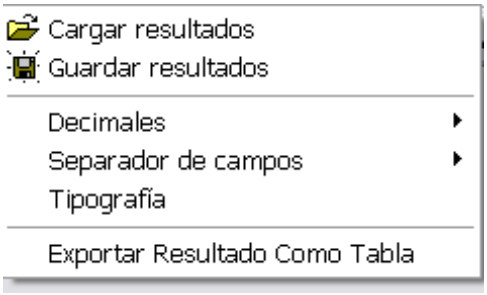
Cruzar categorías

Menú DATOS ⇒ CRUZAR CATEGORIAS permite obtener las combinaciones que resultan al cruzar las categorías de dos o más variables. En la ventana de diálogo bajo la

opción **Criterios de clasificación** se deben indicar las variables a cruzar. Al **Aceptar** aparecerá en la tabla una columna nueva con las clases obtenidas en el cruzamiento.

Resultados

El menú RESULTADOS muestra las acciones que se pueden aplicar al resultado activo (último resultado de una acción solicitada desde el menú Estadística o el menú Aplicaciones). Para activar otro resultado obtenido con anterioridad haga clic en la solapa que lo indexa y que se encuentra ubicada al pie de la ventana **Resultados**. Al activar el menú **Resultados** se podrá elegir entre alguna de las siguientes opciones:



Cargar resultados

Permite abrir un archivo que contiene resultados que han sido guardados en una sesión de trabajo. En la ventana de diálogo se especifica el nombre del archivo y su ubicación.

Guardar resultados

Permite crear un archivo con resultados que han sido obtenidos en una sesión de trabajo. En la ventana de diálogo se especifica el nombre del archivo y su ubicación. Los archivos tendrán extensión ".ITRES".

Decimales

Este ítem despliega un submenú que permite elegir el número de decimales deseados para la presentación. Al final de este menú aparece activada la opción notación exponencial; en el caso que un resultado no pueda mostrar ningún dígito significativo con el número de decimales especificados, InfoStat utilizará notación exponencial.

Separador de campos

Permite elegir un tipo de separador (espacio, tabulador, coma o punto y coma) como carácter que separa columnas de una tabla, por defecto el separador es espacio. Usualmente, este separador no necesita ser modificado pero puede resultar útil en el contexto de la exportación de resultados como tabla.

Tipografía

Permite cambiar los atributos tipográficos (tipo y tamaño de letra) utilizados en la presentación de resultados. Esta acción también se puede invocar activando el botón "A" de la barra de herramientas.

Exportar resultados como tabla

Permite exportar el texto de la ventana de RESULTADOS como una tabla. Al seleccionar esta acción se abrirá una ventana de diálogo llamada **Importador de Texto**. Para detalles de operación con esta ventana ver el punto ABRIR TABLA del menú DATOS.

Acceso a submenúes de resultados mediante el botón derecho del ratón

Además de las acciones que figuran en el Menú RESULTADOS, teniendo activa una ventana de **Resultados** al presionar el botón derecho del ratón se puede acceder a las siguientes opciones:

Decimales: establece la cantidad de decimales que se muestran en la ventana activa.

Copiar: copia el texto previamente seleccionado, con tabuladores como separación de campos. El mismo puede ser leído directamente en procesadores de texto para la construcción de tablas.

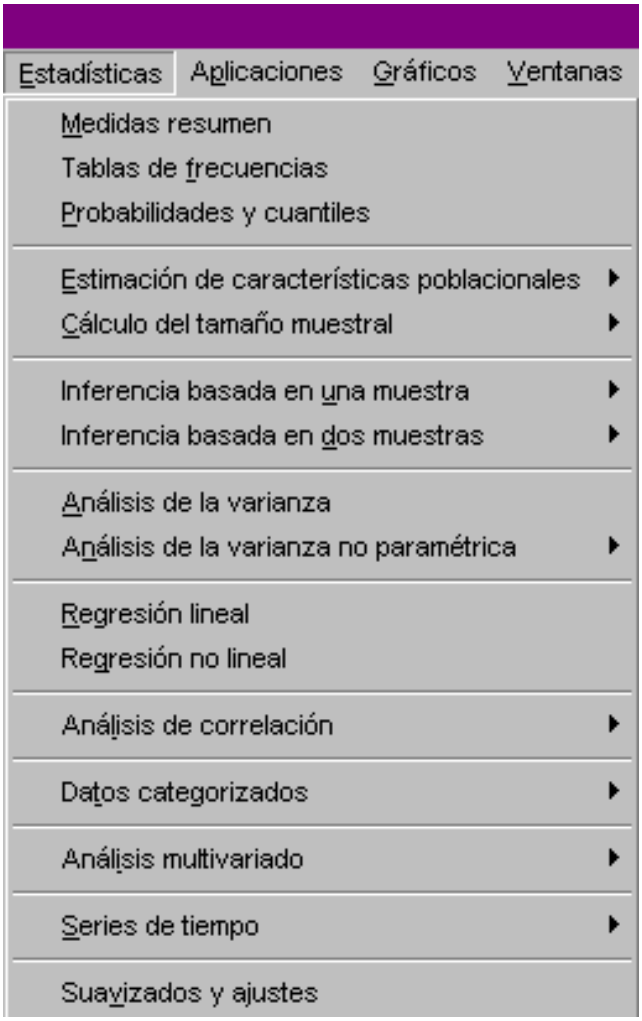
Borrar: borra el resultado activo.

Borrar éste y anteriores: borra el resultado activo y todos los anteriores.

Imprimir: Imprime el contenido del resultado activo.

Estadísticas

InfoStat realiza diversos análisis estadísticos sobre una tabla de datos activa. La selección del tipo de análisis se realiza desde el menú ESTADÍSTICAS. Cada vez que un procedimiento es invocado, las salidas son presentadas en una ventana de resultados, la cual puede ser formateada y preparada para exportación siguiendo especificaciones dadas por el usuario desde el menú RESULTADOS.



Las acciones (submenúes) que se aplican al análisis de tablas InfoStat, en el menú ESTADÍSTICAS, son: **Medidas resumen**, **Tablas de frecuencias**, **Probabilidades y cuantiles**, **Estimación de características poblacionales**, **Cálculo del tamaño muestral**, **Inferencia basada en una muestra**, **Inferencia basada en dos muestras**, **Análisis de la varianza, Análisis de la varianza no paramétrica**, **Regresión lineal**, **Regresión no lineal**, **Análisis de correlación, Datos categorizados, Análisis multivariado, Series de Tiempo, Suavizados y Ajustes**.

En general, estas acciones invocan en primera instancia una ventana que sirve para seleccionar variables. En ella se debe indicar la o las variables de interés y la partición deseada, en caso que el análisis sea por grupo o partición del archivo de datos. En el selector de variables, haciendo *click* sobre las variables de interés, se las incorpora con las flechas a la subventana **Variables**. Las variables que producen particiones en el archivo de datos se deben declarar en la solapa **Particiones**, el comando **Seleccionar por**, permite hacer referencia a la/s variable/s en base a las cuales se quiere diferenciar el análisis. Cuando se selecciona más de

una variable, los grupos resultan de la combinación de los niveles de las variables seleccionadas.

Por ejemplo, si se tienen las variables: color de semilla (claro, oscuro y rojizo) y tamaño de semilla (grande, mediana y chica), al seleccionar sólo el color se crearán tres grupos (los tres niveles de color); si en cambio se eligen ambas variables se tendrán 9 grupos. Las particiones realizadas se visualizan en una lista a la derecha de la ventana que puede ser alterada a través de la selección y eliminación, mediante las flechas de desplazamiento que se encuentran al pie de la lista, de uno o más grupos que no se desea participen en el análisis. Cuando se han identificado grupos, InfoStat realizará el análisis solicitado repetidamente a partir de las observaciones de cada grupo por separado.

Estadística descriptiva

El primer bloque del menú Estadísticas permite describir un conjunto de datos mediante medidas resumen univariadas, tablas de frecuencias y ajustes de funciones de distribución teóricas sobre distribuciones empíricas (tablas de frecuencia muestrales). Todas las acciones pueden realizarse para el conjunto de filas activas como un todo o para cada subgrupo o partición del archivo si se indica una variable que lo particione en la solapa **Particiones**. Para medidas resumen y tablas de frecuencias, es posible trabajar con archivos donde haya tantas filas como observaciones (ver como ejemplo archivo *Atriplex.idb*) o donde cada fila de la columna de interés represente un valor de la variable y exista otra columna del archivo que contiene la frecuencia de cada valor (ver como ejemplo archivo *Insectos.idb*). En el primer caso, en el selector de variable deberá indicarse la o las variables de análisis y no se deberá llenar el campo opcional **Frecuencias**. En el segundo caso, se deberá indicar la columna que contiene los distintos valores de la variable en la ventana **Variables** del selector y la columna que contiene las frecuencias en la ventana **Frecuencias (opcional-solo una)**. InfoStat también ofrece en este bloque un calculador de probabilidades y cuantiles para distintos tipos de variables aleatorias.

Medidas resumen

Se dispone de las siguientes medidas de resumen: número de observaciones **(n)**, **Media**, desviación estándar **(D.E)**, varianza con denominador *n-1* **(Var(n-1))**, varianza con denominador *n* **(Var(n))**, error estándar **(E.E.)**, coeficiente de variación **(CV)**, valor mínimo **(Mín)**, valor máximo **(Máx)**, **Mediana**, cuantil 0.25 o primer cuartil **(Q_1)**, cuantil 0.75 o tercer cuartil **(Q_3)**, suma de las observaciones **(Suma)**, **Asimetría**, **Kurtosis**, suma de cuadrados no corregida **(Suma Cuad.)**, suma de cuadrados corregida por la media **(SCC)**, mediana de los desvíos absolutos respecto de la mediana **(MAD)**, **Datos faltantes**, percentiles 5, 10, 25, 50, 75, 90 y 95 **(P(05), P(10),** etc.).

El *número de observaciones* reportado corresponde al número de casos activos. Los estadísticos muestrales son calculados usando como tamaño de muestra el número de casos

obtenidos después de descartar las observaciones con datos faltantes. El código de datos faltantes puede ser ingresado por el usuario. El estadístico *media* se refiere a la media aritmética. *Desviación estándar* corresponde a la raíz cuadrada de la varianza muestral calculada como la suma de los cuadrados de los desvíos con respecto a la media muestral, dividida por (n–1). El *error estándar* corresponde al desvío estándar dividido por raíz de n. El *coeficiente de variación* es el cociente entre la desviación estándar y la media muestral, expresado en porcentaje.

El *primer cuartil* (Q_1), la *mediana* y el *tercer cuartil* (Q_3) al igual que cualquier otro percentil pueden ser obtenidos mediante el ordenamiento de la muestra y la selección de uno de los valores observados de acuerdo a su posición o bien estimados a partir de una aproximación de función de distribución empírica. Si el usuario selecciona **Basados FDE** en la subventana **Percentiles**, InfoStat estimará previamente la función de distribución mediante el método propuesto por Collings y Hamilton (1988) y luego usará esta función para reportar el percentil solicitado. Si se elige la opción **Muestrales**, el percentil será uno de los valores de la muestra obtenido después del ordenamiento de la misma. Por este motivo, ambos procedimientos no producen necesariamente el mismo resultado numérico.

La presentación de los resultados puede ser orientada en forma horizontal o vertical. La primera es útil para exportar los resultados en una nueva tabla de datos con el objeto de realizar análisis posteriores sobre una tabla de datos conteniendo las medidas resumen.

Las medidas resumen pueden ser solicitadas para una o más variables del archivo simultáneamente (variables indicadas en el selector de variables). Dichas medidas pueden ser obtenidas utilizando todas las observaciones del archivo o bien para subgrupos de observaciones. Los subgrupos pueden ser formados a partir de una variable del archivo o de la combinación de dos o más variables del mismo. Para la formación de subgrupos el usuario puede indicar las variables que definen los grupos listando las mismas en la subventana **Criterios de clasificación (opcional)** del selector de variables. Alternativamente se puede activar la solapa **Partición** para indicar las variables que particionan el archivo, pero esta última opción es menos eficiente que el uso de **Criterios de clasificación** en cuanto al tiempo de ejecución. Por ello es recomendado utilizar criterios de clasificación cuando se desean obtener medidas resumen para un gran número de subgrupos de un archivo extenso.

A modo de ejemplo se usaron los datos del archivo *Atriplex*. Eligiendo el Menú ESTADÍSTICAS ⇒ MEDIDAS RESUMEN, se habilitó la ventana **Estadística descriptiva** en la que se seleccionan la o las variables que se desean analizar, si se especifica una variable para producir una partición del archivo en la solapa **Partición**, se obtendrán las medidas resumen solicitadas para cada grupo o partición. En este ejemplo se seleccionaron las variables "PG", "PN" y "PS", y en la solapa **Partición** se indicó la variable "tamaño". Se activaron o solicitaron las siguientes medidas: **n**, **Media**, **D.E.**, **Var(n-1)**, **Mín**, **Máx**, **Mediana** y **P(50)** estimado a partir de la función de distribución empírica (este estadístico no coincide exactamente con la Mediana ya que la ésta es calculada a partir de los datos muestrales y el P(50) a partir de la *distribución* de los datos muestrales. Si cuando se solicita

el P(50) se deja activada la casilla **Muestrales** entonces la Mediana y P(50) serán iguales). Se seleccionó además **Presentación Horizontal**. Los resultados obtenidos se muestran en la siguiente tabla:

Tabla 3: Estadística descriptiva para variables del archivo Atriplex, de acuerdo a la partición realizada por tamaño de semillas (disposición horizontal).

Estadística Descriptiva

Tamaño	Variable	n	Media	D.E.	Var(n-1)	Mín	Máx	Mediana	P(50)
chicas	Germinación	9	54.56	26.34	694.03	20.00	93.00	60.00	48.67
chicas	PN	9	24.44	20.24	409.53	0.00	60.00	20.00	20.00
chicas	PS	9	0.29	0.02	0.00	0.26	0.31	0.30	0.28
grandes	Germinación	9	73.33	19.28	371.75	40.00	93.00	80.00	71.00
grandes	PN	9	51.33	22.12	489.50	27.00	87.00	47.00	42.33
grandes	PS	9	0.39	0.06	0.00	0.30	0.49	0.40	0.39
medianas	Germinación	9	68.78	32.81	1076.19	13.00	100.00	87.00	80.00
medianas	PN	9	50.67	27.44	752.75	7.00	80.00	54.00	40.50
medianas	PS	9	0.34	0.04	0.00	0.30	0.43	0.32	0.32

Tablas de frecuencias

Menú ESTADÍSTICAS ⇒ TABLAS DE FRECUENCIAS, permite obtener una tabla de frecuencias y/o probar el ajuste de modelos distribucionales teóricos sobre una distribución de frecuencia empírica. Las tablas de frecuencias pueden, de acuerdo a los campos activados por el usuario, contener la siguiente información: límites inferiores **(LI)** y superiores **(LS)**, de los intervalos de clase, marca de clase **(MC)**, frecuencias absolutas **(FA)**, frecuencias relativas **(FR)**, frecuencias absolutas acumuladas **(FAA)** y frecuencias relativas acumuladas **(FRA)**. El número de clases, puede ser obtenido en forma automática o definido por el usuario (PERSONALIZADO). Para la forma automática InfoStat obtiene el número de clases tomando el $\log_2(n+1)$. Para el caso personalizado, InfoStat permite especificar el mínimo, máximo y número de intervalos. Los intervalos que construye son cerrados a la derecha. Si la variable es categórica, la personalización no es aceptada y la tabla de frecuencias presentará tantas clases como categorías tenga la variable. Si los valores de la variable fueron declarados como enteros, InfoStat tiene la opción, por defecto, de considerarla como una variable de conteo y muestra las frecuencias de todos los valores enteros entre el mínimo y el máximo. Si la variable contiene valores enteros y se desactiva

la casilla **Tratar a las variables enteras como conteo**, InfoStat tratará a la variable como continua definiendo intervalos de clases y construyendo la tabla a partir de ellos.

Siguiendo con los datos del archivo *Atriplex*, se obtuvo la tabla de frecuencias para la variable germinación para cada una de los tamaños de semilla invocando las siguientes acciones: ESTADÍSTICAS ⇒ TABLAS DE FRECUENCIAS, en la ventana **Distribución de frecuencias-solapa variables** se seleccionó germinación y antes de **Aceptar** se activó la solapa **Particiones...** y en la subventana **Seleccionar por** se pasó la variable tamaño (automáticamente se visualizan todos los tamaños de semilla presentes en el archivo). Al **Aceptar** aparece la ventana **Distribución de Frecuencias-Opciones de la Tabla de Frecuencias** donde el usuario puede indicar el tipo de información que desea visualizar en la tabla y la definición del número de clases. En este caso se aceptaron todas las opciones que se encuentran activadas por defecto, por lo que sólo se presionó **Aceptar** y el número de clases fue calculado automáticamente. Los resultados se muestran en la siguiente tabla:

Tabla 4: Tabla de frecuencias para la variable germinación del archivo Atriplex, de acuerdo a la partición realizada por la variable tamaño de semillas.

Tabla de Distribucion de Frecuencias

Tamaño	Variable	Clase	LI	LS	MC	FA	FR
chicas	Germinacion	1	20.00	44.33	32.17	3	0.33
chicas	Germinacion	2	44.33	68.67	56.50	3	0.33
chicas	Germinacion	3	68.67	93.00	80.83	3	0.33

Tamaño	Variable	Clase	LI	LS	MC	FA	FR
grandes	Germinacion	1	40.00	57.67	48.83	3	0.33
grandes	Germinacion	2	57.67	75.33	66.50	0	0.00
grandes	Germinacion	3	75.33	93.00	84.17	6	0.67

Tamaño	Variable	Clase	LI	LS	MC	FA	FR_
medianas	Germinacion	1	13.00	42.00	27.50	3	0.33
medianas	Germinacion	2	42.00	71.00	56.50	0	0.00
medianas	Germinacion	3	71.00	100.00	85.50	6	0.67

Ajustes

Menú ESTADÍSTICAS ⇒ TABLAS DE FRECUENCIAS, solapa **Ajustes**, permite obtener pruebas de bondad de ajuste. La hipótesis nula especifica un modelo distribucional teórico para los datos. Los valores observados en la muestra son comparados con los valores esperados según el modelo especificado, mediante el uso del estadístico **Chi cuadrado** y/o el estadístico máximo verosímil **G** (Agresti, 1990). El usuario deberá seleccionar entre uno de estos dos estadísticos para realizar la prueba de bondad de ajuste. Además deberá especificar si desea **estimar** desde la muestra o **especificar** externamente los parámetros de la distribución teórica que hipotéticamente tienen los datos. Si se activa **especificar** aparecerán tantas casillas como parámetros tenga la distribución teórica seleccionada, para recibir la información desde el usuario. Las casillas reservadas para cada parámetro de una distribución contendrán automáticamente los valores de los estimadores muestrales de los mismos. En caso de variables continuas, la distribución empírica se construirá a partir de la información sobre intervalos de clase automáticamente generados. Estos intervalos pueden

ser generados con **límites inferiores y superiores abiertos o cerrados** según especifique el usuario en la ventana **Distribución de Frecuencias-Ajustes**.

Las distribuciones teóricas que se pueden especificar automáticamente en la hipótesis nula son: **Normal**, **Chi cuadrado (Chi Cuad.)**, **Uniforme**, **Binomial, Poisson** y **Binomial negativa (BinNeg)** (ver Capítulo Manejo de Datos). La opción **Ninguna** (seleccionada por defecto) permite visualizar la función de distribución empírica.

Ejemplo 2: Los datos del archivo Insectos, muestran las frecuencias observadas del número de insectos por planta en un lote de 200 plantas. Estos valores se usan para probar la hipótesis que la distribución de la variable ajusta el modelo binomial negativo.

En la ventana **Tabla de distribución de frecuencias**, subventana **Variable** ingrese "Insectos" y en **Frecuencias** ingrese "observados". En **Ajustes** elegir **Binomial negativa**. Se obtendrá una tabla conteniendo las frecuencias absolutas observadas (**FA**), las frecuencias absolutas esperadas de acuerdo con el modelo distribucional propuesto (**E(FA)**), y el **valor *p*** de la prueba de bondad de ajuste.

Tabla 5: Prueba de bondad de ajuste (estadístico Chi cuadrado Pearson) para la hipótesis que las observaciones provienen de una distribución Binomial Negativa con parámetros estimados a partir de la muestra. Archivo Insectos.

Tabla de Distribución de Frecuencias
Ajuste: Binomial Negativa con estimación de parámetros: k=1.10915 y media=1.205000

Variable	Clase	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado	p
Insectos	1	0	89	0.45	88.46	0.44	3.2E-03	
Insectos	2	1	52	0.26	51.09	0.26	0.02	
Insectos	3	2	24	0.12	28.06	0.14	0.61	
Insectos	4	3	15	0.08	15.14	0.08	0.61	
Insectos	5	4	10	0.05	8.10	0.04	1.05	
Insectos	6	5	5	0.03	4.31	0.02	1.16	
Insectos	7	6	4	0.02	2.28	0.01	2.45	
Insectos	8	7	1	0.01	1.21	0.01	2.49	
Insectos	9	8	0	0.00	1.35	0.01	3.83	0.6991

Un valor *p* menor al nivel de significación nominal de la prueba conduce al rechazo del modelo distribucional propuesto. En este ejemplo se puede decir que la distribución del conteo de insectos puede modelarse con la distribución Binomial negativa con los parámetros especificados en el encabezamiento de la tabla dado que $p>0.05$. Los parámetros son estimados automáticamente a partir de la muestra en estudio.

Probabilidades y cuantiles

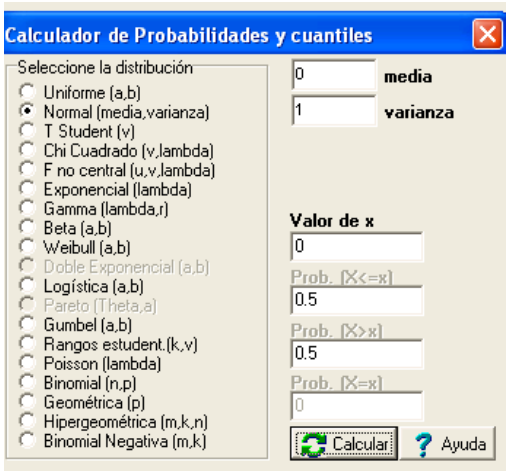
InfoStat provee un calculador para obtener la probabilidad de valores menores o iguales que un valor especificado (Probabilidades Acumuladas), para una amplia lista de variables aleatorias. Los cálculos de probabilidades pueden realizarse bajo los siguientes modelos distribucionales: **Uniforme (a,b)**, **Normal (m,v)**, **T de Student (v)**, **Chi-cuadrado (v,lambda)**, **F no central (u,v,lambda)**, **Exponencial (lambda)**, **Gamma (lambda,r)**,

Beta (a,b), **Weibull (a,b)**, **Logística (a,b), Gumbel (a,b)**, **Rangos estudentizados (k,v)**, **Poisson (lambda), Binomial (n,p)**, **Geométrica (p)**, **Hipergeométrica (m,k,n)** y **Binomial negativa (m,k)** (ver Capítulo Manejo de Datos). Para cada modelo deben especificarse el o los valores de sus parámetros, cuya notación se encuentra entre paréntesis al lado del nombre de la distribución.

InfoStat también provee cuantiles distribucionales bajo estos modelos.

Para obtener un valor de **Probabilidad** primero seleccione la distribución teórica sobre la que desea calcular probabilidades, luego ingrese los parámetros que la caracterizan. Por ejemplo, para el caso de la distribución normal ingrese la media (m) y la varianza (v) de la distribución.

Si desea conocer la probabilidad acumulada hasta un cierto valor (x) de esa distribución active el casillero **Valor de x** ingresando el valor de la variable aleatoria para el cual desea obtener la probabilidad acumulada. Presionando el botón **Calcular** o accionando la tecla Enter, podrá leer en el casillero **Prob. (X≤ x)** la probabilidad de ocurrencia de valores menores o iguales al valor *x*, bajo el modelo distribucional propuesto. En el casillero **Prob. (X>x)** se mostrará el complemento de **Prob. (X≤ x)**. En el casillero **Prob. (X=x)** se muestra la probabilidad de que una variable discreta asuma valores iguales a *x*, bajo el modelo distribucional propuesto (si se selecciona un modelo para variables continuas, este valor será siempre cero). Si desea conocer el cuantil *p* de la distribución seleccionada, ingrese el valor *p* en el casillero **Prob. (X≤ x)** y presione **Calcular**. Se podrá leer en el casillero **Valor de x** el cuantil p-ésimo del modelo distribucional propuesto, donde $p \in [0,1]$.



Estimadores de características poblacionales

Este módulo permite estimar características poblacionales en estudios muestrales diseñados bajo las siguientes técnicas: muestreo aleatorio simple, muestreo estratificado y muestreo por conglomerados.

Definiciones de términos relacionados al muestreo

Una **población** (o universo) es un conjunto de elementos o entidades que comparten algún atributo y cuyos límites temporales o espaciales pueden establecerse. Las poblaciones pueden ser finitas o infinitas según su tamaño. Las poblaciones finitas tienen una cantidad numerable de objetos. El **elemento o unidad elemental** es un objeto o individuo de la

población sobre el cual se toma efectivamente la lectura o medición de la característica en estudio. Una **muestra** es todo subconjunto no vacío de la población que simbolizaremos por $\{X_1,X_2,...,X_n\}$. No toda muestra es adecuada y pertinente para los objetivos de un estudio, de allí la necesidad de diseñar el esquema de muestreo y obtener estimaciones de acuerdo a la técnica utilizada en la recolección de información. Los elementos o conjuntos de ellos que son objeto de selección por un proceso de muestreo se conocen como **unidades muestrales**. El conjunto total de unidades muestrales en una población se define como el **marco muestral**.

Por ejemplo, se desea conocer el nivel de infestación por mosca del mediterráneo de los frutos de una plantación de duraznos. La **población** es la colección de todos los duraznos en la plantación. El **elemento muestral** es el durazno. Puede ser dificultoso construir el marco muestral a partir de los duraznos individuales, pero se podría hacer a partir de cada planta, luego las **unidades muestrales** serían las plantas de durazno. El **marco muestral** es el conjunto de todas las plantas en la plantación objeto de estudio.

Los **parámetros** son constantes que caracterizan una población, como por ejemplo la media poblacional, la proporción de casos con un atributo dado, el total de un atributo y la varianza poblacional. Los **estimadores** son funciones definidas sobre el espacio de todas las muestras posibles de un tamaño dado y sus imágenes tienen por objeto proveer información sobre el valor de los parámetros poblacionales. Ejemplo de estimadores son la *media* y la *varianza muestral*.

InfoStat admite dos tipos de variables para producir estimaciones de parámetros poblacionales. Las características en estudio pueden ser continuas o dicotómicas. Características dicotómicas permiten estimar parámetros poblacionales relacionados a la proporción de éxitos o casos en una clase determinada. Si el usuario desea convertir una variable continua en otra dicotómica, InfoStat permite dicotomizar variables a partir de la comparación de cada uno de sus valores con un valor de referencia. El punto que permite la dicotomización puede ser la media de la característica, la mediana o un valor arbitrario ingresado por el usuario. Se puede dicotomizar denominando "éxito" a los valores de la variable en estudio mayores, menores, mayores o iguales, menores o iguales que un valor de referencia ingresado por el usuario.

Sea $\{X_1, X_2 ,..., X_N \}$ el conjunto de todos los valores en la población (población de tamaño N), entonces definimos los parámetros total, media y varianza como:

$$\textit{Total} \qquad \tau = \sum_{i=1}^{N} X_i$$

$$\textit{Media} \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\textit{Varianza} \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2$$

En una población de tamaño *N*, el número de muestras posibles de tamaño *n*, con un muestreo sin reposición es C(*N*,*n*) (combinatorio de *N* tomados de a *n*). Ejemplo si *N*=30 y *n*=2 C(30,2)=435. Si se calcula un estadístico muestral a partir de cada muestra tendremos 435 estimadores muestrales eventualmente diferentes. Esto origina lo que llamamos la distribución muestral del estimador. El **error estándar** de un estimador corresponde a la raíz cuadrada de la varianza de dicha distribución muestral. El **coeficiente de variación** de un estimador de un parámetro poblacional se define como el cociente entre su error estándar y el verdadero valor del parámetro estimado. El cuadrado del coeficiente de variación de un parámetro estimado es referido como la **varianza relativa** del parámetro estudiado. El error estándar de un estimador es una medida de la variabilidad muestral del estimador sobre todas las muestras posibles. Si se asume que la distribución de los estimadores se aproxima, cuando el tamaño muestral es suficientemente grande, a la distribución normal, es posible utilizar la teoría normal para obtener intervalos de confianza aproximados para los parámetros que están siendo estimados. El intervalo de confianza (1-α)% para el parámetro θ será:

$$\hat{\theta} \pm Z_{1-\frac{\alpha}{2}} EE\left(\hat{\theta}\right)$$

siendo $\hat{\theta}$ el estimador de θ; $Z_{1-\frac{\alpha}{2}}$ el percentil $\left(1-\frac{\alpha}{2}\right)100$ de la distribución normal estándar y $EE\left(\theta\right)$ el error estándar de $\hat{\theta}$.

Para los distintos tipos de muestreo y estimadores disponibles, InfoStat permite obtener el error estándar, el coeficiente de variación del estimador, la varianza relativa y el intervalo de confianza para los parámetros estimados con el coeficiente de confianza requerido por el usuario.

Muestreo aleatorio simple

Menú ESTADÍSTICAS ⇒ ESTIMACIÓN DE CARACTERÍSTICAS POBLACIONALES ⇒ MUESTREO ALEATORIO SIMPLE, permite estimar parámetros poblacionales en el marco de un muestreo aleatorio simple. El *muestreo aleatorio simple* (m.a.s), es un plan de muestreo en el que se toma una muestra de tamaño *n*, con un procedimiento tal que, toda muestra de tamaño *n* (de una población de *N* elementos) tiene la misma probabilidad de ser elegida. El número total de muestras posibles es *T*=C(*N*,*n*). La probabilidad de elección de una muestra m_j de tamaño *n* es:

$$P(m_j)=1/T \quad \text{con } j=1,...,T$$

InfoStat supone que los valores en las columnas de la tabla de datos corresponden a valores muestrales de una o más características en estudio. En la ventana de diálogo del selector de variables se debe indicar cual o cuales son las columnas de la tabla que contienen estas características. Cuando existen diversos criterios de clasificación en la población, pero por consideraciones teórico-prácticas no es conveniente realizar un muestreo estratificado se

pueden realizar estimaciones dentro de estos **subdominios** a través de un muestreo aleatorio simple. La población puede ser finita, y en tal caso hay que ingresar el tamaño poblacional.

Por conveniencia se denotará a los elementos muestrales del primero al enésimo con $x_1,...,x_n$. Luego estos son los valores de la variable *X* para los elementos 1 al *n*. Después de haber tomado la muestra, es posible calcular valores como: *totales*, *medias*, *proporciones*, *desvíos estándares*, etc.

InfoStat estima, bajo muestreo aleatorio simple, el total, la media y la proporción de éxitos (y total de éxitos), de la siguiente manera:

$$t_{mas} = \frac{N}{n}\sum_{i=1}^{n} x_i$$

$$\overline{X}_{mas} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$p_{mas} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(x_i)$$

con $\mathrm{I}(x_i)$ función indicadora que evaluada en la observación x_i devuelve un "1" o "0" conforme la observación represente u éxito o un fracaso respectivamente.

Se pueden requerir intervalos de confianza para los parámetros poblacionales con un nivel de confianza especificado por el usuario. Por defecto el intervalo que se construye tiene un coeficiente de confianza del 95%. Para construir dichos intervalos se utilizan los errores estándares de los estimadores correspondientes, los cuales son calculados como la raíz cuadrada de las siguientes varianzas,

$$V(t_{mas}) = N^2\,\frac{N-n}{N}\,\frac{S_X^2}{n}$$

$$V(\overline{X}_{mas}) = \frac{N-n}{N}\,\frac{S_X^2}{n}$$

$$V(p_{mas}) = \frac{N-n}{N}\,\frac{p(1-p)}{n-1}$$

donde S_X^2 es el estimador insesgado de la varianza poblacional de la característica *X* en estudio, bajo el supuesto de población infinita y se define como:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Los estimadores precedentes involucran un factor de corrección por finitud que es utilizado en caso de poblaciones finitas. Si no se especifica el tamaño de la población, InfoStat asume

población infinita y no utiliza factor de corrección por finitud. También se pueden solicitar el coeficiente de variación y la varianza relativa asociados a la estimación obtenida.

Al invocar este submenú en InfoStat, aparece la ventana **Muestreo aleatorio simple** que permite elegir las variables y particiones deseadas. El criterio **Particiones** de InfoStat puede ser utilizado en este menú, para obtener estimaciones para distintas particiones del archivo, definidas en función de una o más variables. En caso de que existan **subdominos** se deberá indicar a InfoStat cuál es la columna de la tabla de datos que los identifica. Si existe una columna del archivo conteniendo frecuencias absolutas para cada valor de la característica en estudio y dicha columna es indicada en la subventana **Frecuencias**, InfoStat usará esa información para ponderar los valores de la característica por su frecuencia para cualquier estimación que se solicite a continuación. Al aceptar se habilita otra ventana que permite **Ingresar el tamaño poblacional**. La opción **Características continuas** habilita una subventana **Características poblacionales a estimar** en la que se puede activar: **Promedio**, **Total**, **Proporción de éxitos** y **Total de éxitos**. Cuando se eligen alguna de estas dos últimas opciones se ingresa automáticamente a **Dicotomizar por** y en **Considerar éxito los valores** están las siguientes opciones: **mayores que la media**, **mayores o iguales que la media**, **menores que la media**, **menores o iguales que la media**, **mayores que la mediana**, **mayores o iguales que la mediana**, **menores que la mediana**, **menores o iguales que la mediana** y **mayores que**, **mayores o iguales que**, **menores que**, **menores o iguales que** un valor determinado ingresado por el usuario en la ventana dispuesta para tal fin.

En la parte inferior de la ventana principal aparecen las siguientes opciones: Error estándar del estimador, Intervalo de confianza para el parámetro poblacional, Coeficiente de variación del estimador y Varianza relativa.

Muestreo estratificado

Menú ESTADÍSTICAS ⇒ ESTIMACIÓN DE CARACTERÍSTICAS POBLACIONALES ⇒ MUESTREO ESTRATIFICADO, permite obtener estimaciones en el marco de un *muestreo estratificado.* En este tipo de plan de muestreo, la población es dividida en estratos y una muestra aleatoria simple es tomada de cada estrato. Si se denota por N_h al tamaño del estrato h, con n_h al tamaño de la muestra obtenida desde ese estrato (con h=1,...,L), el total de muestras posibles de tamaño n está dado por:

$$T = \binom{N_1}{n_1} \ \ldots \ \binom{N_h}{n_h} \ \ldots \ \binom{N_L}{n_L}$$

donde la $\sum_{h=1}^{L} n_h = n$

Por ejemplo, si se tienen 3 estratos designados como E1, E2 y E3, de tamaño 3, 5 y 4 respectivamente, el número de muestras posibles de tamaño 2, 3 y 2 para los estratos mencionados serán: 3, 10 y 6. Un ejemplo detallando todas las muestras posibles para la conformación de los estratos mencionados se presenta a continuación:

Población

Estrato	X
1	10
1	11
1	9
2	12
2	13
2	11
2	14
2	13
3	17
3	19
3	18
3	20

Muestras posibles

	E1		E2			E3	
M1	10	11	12	13	11	17	19
M2	10	9	12	13	14	17	18
M3	11	9	12	13	13	17	20
M4			12	11	14	19	18
M5			12	11	13	19	20
M6			12	14	13	18	20
M7			13	11	14		
M8			13	11	13		
M9			13	14	13		
M10			11	14	13		

Los estimadores por estrato (indexado por *h*) del total, media y proporción poblacional son:

$$t_h = \frac{N_h}{n_h}\sum_{i=1}^{n_h} x_{ih}$$

$$\bar{X}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{ih}$$

$$p_h = \frac{1}{n_h}\sum_{i=1}^{n_h} \mathrm{I}(x_{ih})$$

donde x_{ih} es el i-ésimo valor de la variable observada en el estrato *h* y $I_{(xih)}$ es una función indicadora que evaluada en la observación X_i devuelve un "1" o un "0" conforme la observación representa un éxito o un fracaso respectivamente.

Estos estimadores tienen la misma forma que los estimadores para muestreo aleatorio simple dentro de cada estrato. Por ende la varianza de los estimadores promedio (me) a través de *L* estratos se construye a partir de las varianzas de los estimadores por estrato.

$$V\left(t_{me}\right)=\sum_{h=1}^{L}N_h^2\,\frac{S_h^2}{n_h}\left(\frac{N_h-n_h}{N_h}\right)$$

$$V\left(\bar{X}_{me}\right)=\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\,\frac{S_h^2}{n_h}\left(\frac{N_h-n_h}{N_h}\right)$$

$$V\left(p_{me}\right)=\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\,\frac{p_h\left(1-p_h\right)}{n_h-1}\left(\frac{N_h-n_h}{N_h}\right)$$

donde S_h^2 es la varianza de la variable aleatoria en el estrato *h*.

En algunas circunstancias, las unidades muestrales no pueden clasificarse *a priori* como pertenecientes a un estrato dado. Si esa información es obtenida en el propio proceso de muestreo, se usa entonces un muestreo *post-estratificado.* Esta técnica se basa en un muestreo aleatorio simple a partir del cual se aplican los estimadores para muestreo estratificado, previa clasificación de las unidades muestrales en los distintos estratos. La diferencia con la estimación para sudominios en el marco del muestreo aleatorio simple es que los tamaños de los estratos en este caso son conocidos. La varianza de los estimadores es corregida para tener en cuenta que los tamaños muestrales resultantes por estrato sean aleatorios.

Al invocar el submenú MUESTREO ESTRATIFICADO en InfoStat, aparece la ventana **Muestreo estratificado** que permite seleccionar las variables a usar. El criterio partición de InfoStat puede ser utilizado en este menú, para obtener estimaciones para distintas particiones del archivo, definidas en función de una o más variables. En este caso la declaración del **Estrato** es obligatoria. Si existe una columna del archivo conteniendo frecuencias absolutas para cada valor de la característica en estudio y dicha columna es indicada en la subventana **Frecuencias**, Infostat usará esa información para ponderar los valores de la característica por su frecuencia para cualquier estimación que se solicite a continuación. Al aceptar se habilita otra ventana donde hay una **Lista de los nombres de los estratos**, en la que se debe ingresar el **Tamaño de los estratos.** Si el muestreo es post-estratificado se debe activar el campo correspondiente.

La opción **Características continuas** habilita una subventana **Características poblacionales a estimar** en la que se puede activar: **Promedio**, **Total**, **Proporción de éxitos** y **Total de éxitos**. Cuando se eligen alguna de estas dos últimas opciones se ingresa automáticamente a **Dicotomizar por** y en **Considerar éxito los valores** están las siguientes opciones: **mayores que la media**, **mayores o iguales que la media**, **menores que la media**, **menores o iguales que la media**, **mayores que la mediana**, **mayores o iguales que la mediana**, **menores que la mediana**, **menores o iguales que la mediana** y **mayores que**, **mayores o iguales que**, **menores que**, **menores o iguales que** un valor determinado por el usuario en la ventana para tal fin.

En la parte inferior de la ventana aparecen las siguientes opciones: Error estándar del estimador, Intervalo de confianza, Coeficiente de variación del estimador y Varianza relativa.

Muestreo por conglomerados

Menú ESTADÍSTICAS ⇒ ESTIMACIÓN DE CARACTERÍSTICAS POBLACIONALES ⇒ MUESTREO POR CONGLOMERADOS, permite obtener estimaciones de parámetros en el marco de un *muestreo por conglomerado.* Este tipo de muestreo se utiliza cuando no es posible o es impráctico contar con un marco muestral de las unidades muestrales elementales y se puede, en cambio, obtener un marco muestral de grupos (conglomerados) de unidades muestrales.

Por ejemplo, si se quiere estimar el grado de ataque de mosca del Mediterráneo en plantas de durazno y el monte cuenta con 20 plantas, cada planta se podría considerar un conglomerado. De estos conglomerados se seleccionan *m* al azar y en cada uno de ellos, se cuenta para cada una de las ramas principales el número de frutos sanos y número de frutos enfermos.

Existen diversos planes de muestreo que genéricamente se engloban bajo el título muestreo por conglomerado, pero cada uno de ellos genera estimadores y errores diferentes. InfoStat realiza las estimaciones correspondientes a un muestreo por conglomerados simple en una etapa. El muestreo por conglomerados simple en una etapa se caracteriza por la elección, según un plan de muestreo aleatorio simple, de un conjunto de *m* conglomerados. Luego estos conglomerados son censados.

La notación utilizada en el marco de este muestreo es la siguiente:

M=número de conglomerados en la población

m=número de conglomerados muestreados

n_c=número de unidades en el conglomerado

N=tamaño de la población

$\bar{N}$ =tamaño promedio de los conglomerados

Los estimadores bajo este esquema de muestreo para características continuas son:

$$t = \left(\frac{M}{m}\right)\sum_{j=1}^{m}\sum_{i=1}^{n_c} x_{ij} = \text{total en la población}$$

$$\bar{X} = \left(\frac{M}{Nm}\right)\sum_{j=1}^{m}\sum_{i=1}^{n_c} x_{ij} = \text{media en la población}$$

$$t_c = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_c} x_{ij}}{m} = \text{total por conglomerado}$$

$$\overline{X}_c = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_c} x_{ij}}{m\overline{N}} = \text{media por conglomerado}$$

Para características binarias, InfoStat permite estimar la proporción de éxitos y el total de éxitos. Cuando la variable es continua se puede calcular la proporción de éxitos y total de éxitos previa dicotomización de la característica continua.

Las varianzas de los estimadores son obtenidas como sigue:

$$V(t) = \frac{M^2}{m}\frac{\sum_{j=1}^{m}\left(\sum_{i=1}^{n_c} x_{ij} - \overline{X}\right)^2}{m-1}\frac{M-m}{M}$$

$$V(\overline{X}) = \frac{M^2}{m}\frac{\sum_{j=1}^{m}\left(\sum_{i=1}^{n_c} x_{ij} - \overline{X}\right)^2}{m-1}\frac{M-m}{M}\frac{1}{N^2}$$

$$V(t_c) = \frac{1}{m}\frac{\sum_{j=1}^{m}\left(\sum_{i=1}^{n_c} x_{ij} - t_c\right)^2}{m-1}\frac{M-m}{M}$$

$$V(\overline{X}_c) = \frac{1}{m\overline{N}^2}\frac{\sum_{j=1}^{m}\left(\sum_{i=1}^{n_c} x_{ij} - t_c\right)^2}{m-1}\frac{M-m}{M}$$

Al invocar este submenú en InfoStat, aparece la ventana **Muestreo por conglomerados** cuya función es permitir elegir las variables y particiones deseadas. El criterio partición de InfoStat puede ser utilizado en este menú, para obtener estimaciones para distintas particiones del archivo, definidas en función de una o más variables. En este caso la declaración del **Conglomerado** es obligatoria (indicar la columna de la tabla de datos que los identifica). Si existe una columna del archivo conteniendo frecuencias absolutas para cada valor de la característica en estudio y dicha columna es indicada en la subventana **Frecuencias**, Infostat usará esa información para ponderar los valores de la característica por su frecuencia para cualquier estimación que se solicite a continuación. Al aceptar se habilita otra ventana donde se debe ingresar **el Número de conglomerados en la población (*M*)** y el **Tamaño promedio de los conglomerados (*N*)**. La opción **Características poblacionales a estimar** permite activar: **Promedio**, **Total**, **Proporción de éxitos** y **Total**

de éxitos. Cuando se desea dicotomizar variables continuas ir a **Dicotomizar por** y en **Considerar éxito los valores** seleccionar alguna de las siguientes opciones: **mayores que la media**, **mayores o iguales que la media**, **menores que la media**, **menores o iguales que la media**, **mayores que la mediana**, **mayores o iguales que la mediana**, **menores que la mediana**, **menores o iguales que la mediana** y **mayores que**, **mayores o iguales que**, **menores que**, **menores o iguales que** un valor determinado por el usuario.

En la parte inferior de esta ventana aparecen marcadas las siguientes opciones: **Error estándar del estimador**, **Intervalo de confianza**, **Coeficiente de variación del estimador** y **Varianza relativa**.

Cálculo del tamaño muestral

Menú ESTADÍSTICAS ⇒ CÁLCULO DE TAMAÑO DE MUESTRA, permite calcular el tamaño de muestra necesario para estimar una media o una proporción poblacional con una confianza y precisión determinada por el usuario. También, permite calcular tamaño de muestra para detectar, en el contexto del ANAVA de efectos fijos a una vía de clasificación, una diferencia entre medias de grupos o poblaciones tan pequeña como sea especificada por el usuario y el tamaño de muestra para la estimación de la diferencia entre dos poblñaciones. Al ingresar a este submenú se habilita la ventana **Tamaño muestral para**... la cual presenta tres solapas: **Detectar una DMS, Estimar una media**,.**Estimar una proporción** y.**Dif. Prop**

Estimar una media

Este método presupone un *m.a.s.* (muestreo aleatorio simple) y tiene por objeto dar una aproximación, basada en la distribución normal, del tamaño muestral necesario para estimar a la media con una confianza y una precisión determinada. La aproximación usada para el cálculo del tamaño de muestra en InfoStat es:

$$n \geq \left(\frac{2Z_{1-\frac{\alpha}{2}}\sigma}{c} \right)^2$$

donde σ es la desviación estándar poblacional, para la que se debe ingresar el valor o una cota superior, *c* es la amplitud requerida para el intervalo de confianza con una confianza (1-α)% para la media poblacional. El valor *c* puede elegirse arbitrariamente o expresarse como una fracción *f* de la media muestral $\left(c = \bar{x}f\right)$.

Alternativamente el usuario puede especificar el error estándar máximo aceptable para la estimación, como criterio para el cálculo del tamaño muestral.

Para detectar una diferencia mínima significativa

Para un diseño balanceado con *a* tratamientos o poblaciones bajo estudio (modelo de efectos fijos), InfoStat provee los tamaños de muestras asociados a valores de potencia, para la prueba de efectos de tratamientos nulos, especificados por el usuario. Los tamaños muestrales por tratamiento son derivados a partir de la relación entre $\phi^2 = \frac{n\sum_{i=1}^{a}\sigma_i^2}{a\sigma^2}$ y la potencia dada por $P(F_0 > F_{\alpha,a-1,N-a} / H_0$ es falsa), donde τ_i es el efecto del tratamiento i-ésimo, σ^2 la varianza común dentro de los tratamientos, *a* el número de tratamientos, α el nivel de significación de la prueba de efectos de tratamientos nulos, *N* el número total de observaciones y F_0 el estadístico del Análisis de la Varianza.

Para evitar que el usuario deba seleccionar el conjunto de τ_i, i=1,...,a, el cálculo se basa en la expresión $\phi^2 = \frac{nD^2}{2a\sigma^2}$ donde D es la mínima diferencia que se quiere detectar entre dos medias.

Si la diferencia entre dos medias es a lo sumo D, el valor de $\phi^2 = \frac{nD^2}{2a\sigma^2}$ y consecuentemente el tamaño de muestra que se obtiene es conservador, esto es, proporciona una potencia al menos igual a la especificada por el usuario.

En la subventana **Criterio para la obtención del tamaño muestral**, se pueden especificar dos alternativas: **Amplitud intervalo de confianza** o **Error estándar de la estimación**. En la medida que se cambien las opciones para estas dos alternativas, en la parte inferior aparecerá un espacio para poner la **Cota superior para la varianza** y así se estimará el **Tamaño muestral requerido**.

Detectar una DMS (diferencia mínima significativa), permite calcular la **Potencia alcanzada** para un modelo de análisis de la varianza de efectos fijos, cuando se van cambiando las siguientes opciones: **Número de tratamientos**, **Varianza común dentro de tratamientos**, **Nivel de significación**, **Mínima diferencia que se quiere detectar** y **Repeticiones por tratamiento (n)**.

Estimar una proporción

Presupone un *m.a.s.* (muestreo aleatorio simple) y tiene por objeto dar una aproximación, basada en la distribución normal, del tamaño muestral necesario para estimar una proporción con una confianza y una precisión determinada. La aproximación usada para el cálculo del tamaño de muestra en InfoStat es:

$$n \geq \left(\frac{2Z_{1-\frac{\alpha}{2}} p(1-p)}{c} \right)^2$$

donde p es la proporción poblacional supuesta a priori, para la que se debe ingresar el valor a través de una barra de movimiento en el rango 0 a 1, *c* es la amplitud requerida para el intervalo de confianza, expresada como un porcentaje de p, con una confianza (1-α)% para la verdadera proporciónal.

Para la estimación de la diferencia entre dos proporciones

En el contexto de un muestreo aleatorio simple, donde se desea estimar la diferencia entre dos proporciones a partir de muestras de igual tamaño, InfoStat provee el tamaño de muestra a extraer desde cada población y los valores asociados de potencia para la prueba de hipotesis de no diferencias de proporciones. Los cálculos son realizados a partir de la aproximación normal (ver estimación de diferencia entre proporciones).

Inferencia en una y dos poblaciones

InfoStat permite contrastar hipótesis y obtener intervalos de confianza para parámetros de un modelo estadístico involucrando una o dos poblaciones. Los menúes de este módulo permiten indicar si la inferencia se basa en una o dos muestras aleatorias. Las acciones (submenúes), que se pueden invocar en el caso de una muestra son: **Prueba T para un parámetro**, **Prueba de rachas**, **Intervalos de confianza**, **Bondad de Ajuste (Kolmogorov)** y **Prueba de normalidad (Shapiro-Wilks modificado)**. En el caso de dos muestras: **Prueba T (muestras independientes)**, **Prueba de Wilcoxon (Mann-Whitney U)**, **Prueba de Wald-Wolfowitz**, **Prueba de Van der Waerden (puntuación normal)**, **Prueba de Bell-Doksum (puntuación normal)**, **Prueba de Kolmogorov-Smirnov**, **Prueba de Irwin-Fisher**, **Prueba de la mediana**, **Prueba para la diferencia de proporciones**, **Prueba T (observaciones apareadas)**, **Prueba de Wilcoxon (observaciones apareadas)**, **Prueba del signo** y **Prueba F para igualdad de varianzas.**

En caso de solicitar el análisis para más de una variable respuesta, los resultados se informan para cada variable por separado.

Inferencia basada en una muestra

Prueba T para un parámetro

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN UNA MUESTRA ⇒ PRUEBA T PARA UN PARÁMETRO, permite probar una hipótesis acerca de la esperanza de una variable aleatoria, del tipo H_0: $\mu=\mu_0$. La prueba utiliza una estimación de la varianza de la variable respuesta.

InfoStat provee el valor *p* para una prueba bilateral, p(Bilateral), o el valor *p* para pruebas unilaterales derecha, p(Unilateral D), o izquierda, p(Unilateral I), según se especifique. Cuando el valor *p* es ≤ que el nivel de significación nominal (α seleccionado para la

prueba), el estadístico pertenece a la región de rechazo, es decir la prueba sugiere el rechazo de la hipótesis nula.

El estadístico de la prueba es: $T = \left(\frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \right)$ que bajo H_0 tiene distribución 'T de Student' con n-1 grados de libertad.

En InfoStat al activar el submenú PRUEBA T PARA UN PARÁMETRO, aparece una ventana con el mismo nombre que permite elegir la variable en estudio y si se desea las variables que definen particiones. La ventana siguiente permite solicitar la información a mostrar y elegir el tipo de prueba a realizar: **Bilateral**, **Unilateral derecha** o **Unilateral izquierda**. Por defecto, InfoStat mostrará la siguiente información: **n** (tamaño muestral), **Media**, **DE** (desviación estándar), **T** (valor del estadístico) y **p** (valor *p*) y el **intervalo de confianza** (por defecto la confianza es del 95% pero se puede optar por otro valor activando el campo correspondiente). El campo **parámetro** permite introducir por teclado el valor hipotetizado para la media poblacional, es decir μ_0.

Siguiendo con los datos del archivo *Atriplex*, se presentan los resultados de la prueba acerca de la media del porcentaje de germinación. Suponga que se desea probar la hipótesis H_0: μ=50. Luego, ingresando el valor 50 en el campo **Parámetro** y simplemente aceptando las opciones activadas, se obtuvieron los siguientes resultados (el análisis se realizó dos veces, una usando una partición del archivo por tamaño de semillas y la segunda sin partición.

Como puede observarse, el porcentaje de germinación es significativamente distinto de 50% sólo para las semillas grandes. La media de germinación sugiere que las semillas de mayor tamaño tienen un porcentaje de germinación mayor al 50%. Trabajando con todos los datos, sin particionar por tamaño, también se rechaza la hipótesis nula.

Tabla 6: Resultados prueba T para datos particionados por tamaño de semillas. Archivo Atriplex.

Prueba T para un parámetro
Valor del Parametro Probado: 50

Tamaño	Variable	n	Media	DE	LI(95)	LS(95)	T	p(Bilateral)
chicas	Germin	9	54.56	26.34	52.25	56.86	0.52	0.6180
grandes	Germin	9	73.33	19.28	73.33	73.33	3.63	0.0067
medianas	Germin	9	68.78	32.81	68.78	68.78	1.72	0.1243

Tabla 7: Resultados prueba T para datos sin particionar. Archivo Atriplex.

Prueba T para un parámetro
Valor del Parametro Probado: 50

Variable	n	Media	DE	LI(95)	LS(95)	T	p(Bilateral)
Germinacion	27	65.56	26.93	63.50	67.61	3.00	0.0059

Prueba de rachas

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN UNA MUESTRA ⇒ PRUEBA DE RACHAS, permite probar la hipótesis de un ordenamiento aleatorio contra una alternativa de tendencia (ordenamiento no aleatorio), mediante el uso de rachas.

Una racha es una sucesión de uno o más elementos, que está precedida y/o seguida de elementos diferentes a los que componen la racha. Para variables dicotómicas se identificará una racha cuando exista una secuencia de valores de la variable que pertenecen a una misma categoría. Por ejemplo, si se tiene la siguiente serie: 1 0 0 0 1 1 0 0 1 1, donde hay tres rachas de unos (de largo 1, 2, y 2) y dos rachas de ceros (de largo 3 y 2).

A modo de ejemplo, suponga que se toman medidas diarias de un indicador económico. Se identificará una racha cuando exista un grupo de medidas consecutivas donde cada uno de los valores diarios sea más alto que el del día previo. Aquí la variable no es dicotómica. InfoStat permite generar variables dicotómicas para el análisis de rachas, en estos casos. El usuario puede indicar un valor, como puede ser la mediana, para establecer la nueva serie dicotómica mediante la comparación de cada observación original con dicho valor.

El estadístico *R* se basa en el número de rachas, en el ejemplo presentado *R*=5. Cuando los tamaños muestrales tienden a infinito, Wald y Wolfowitz demuestran que la estandarización del estadístico *R*, tiende a una distribución normal estándar (Lehmann, 1975) y por tanto puede utilizarse la aproximación normal para el cálculo de valores *p.*

InfoStat permite realizar esta prueba activando el submenú PRUEBA DE RACHAS. Al hacerlo aparece una ventana con el mismo nombre que permite seleccionar la variable en estudio y las que definen particiones. Al aceptar aparece otra ventana donde se puede elegir: **La secuencia dada es aleatoria**, **Tiene tendencia respecto de la mediana** (por defecto) y **Tiene tendencia respecto de** (que habilita un ventana para escribir un valor). En **Mostrar la siguiente información** se encuentra: $\mathbf{n_1+n_2}$, $\mathbf{n_1}$, $\mathbf{n_2}$, **rachas**, **E(R)**, **Est Z** y **p(2 colas)**.

Donde $\mathbf{n_1}$ y $\mathbf{n_2}$ son los números de rachas de las clases 1 y 2 de la variable dicotómica en estudio; **rachas** corresponde al estadístico de la prueba; **R** es el número de rachas de una de las clases (la correspondiente a la primera observación del archivo); **E(R)** es la esperanza del estadístico **R** definida como:

$$E(R)=\left(\frac{2n_1n_2}{n_1+n_2}\right)+1$$

Est Z es el valor del estadístico estandarizado:

$$Est\ Z=\frac{R-E(R)}{S}\ \text{ con }\ S=\sqrt{2\ \frac{n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

Al activar **p(2 colas)** se obtiene el valor *p* de la prueba para la hipótesis nula, la cual puede ser: **La secuencia dada es aleatoria**, **Tiene tendencia respecto de la mediana** de la serie o **Tiene tendencia respecto de** un valor que especifica el usuario.

Cuando los valores n_1 y n_2 son menores que 30, InfoStat obtiene los valores *p* exactos a partir de la distribución del estadístico **R**. Si los valores de n_1 y n_2 son mayores que 30 el valor *p* es obtenido a partir del estadístico **Est Z**.

Intervalos de confianza

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN UNA MUESTRA ⇒ INTERVALOS DE CONFIANZA, permite obtener intervalos de confianza paramétricos con coeficientes de confianza especificados por el usuario para los parámetros **Media**, **Mediana**, **Varianza** y **Proporción**. Estos mismos intervalos más el intervalo de confianza para un **Percentil** de la distribución pueden obtenerse de forma no-paramétrica mediante la técnica de remuestreo *Bootstrap* (Efron y Tibshirani, 1993).

Un intervalo de confianza de nivel α es definido como un conjunto de valores del parámetro (intervalo) que con confianza $(1-\alpha)100\%$ incluirían el valor del parámetro en la población, dado la variabilidad en la muestra y la forma de la distribución muestral del estimador.

Los intervalos de confianza paramétricos se construyen a partir de suposiciones sobre la forma de la distribución muestral del estimador (Normal, T de Student, Chi cuadrado, etc.).

Los cuantiles $\alpha/2$ y $(1-\alpha/2)$ de la distribución muestral del estadístico usado para construir el intervalo, son seleccionados para obtener los límites superior e inferior de un intervalo de nivel α alrededor del parámetro. Intervalos construidos por este proceso tienen, por azar, la posibilidad de no incluir el verdadero valor del parámetro (riesgo tipo I), pero se espera que éste evento suceda sólo en $\alpha\times100\%$ de los intervalos obtenidos.

Consideremos el ejemplo de construcción de un intervalo de confianza, con nivel 0.05, alrededor de la **Media** μ de la población. Se conoce por el teorema central del límite que la media muestral, $\bar{X}$, se distribuye aproximadamente normal alrededor de μ con error estándar $\sigma/\sqrt{n}$ para tamaños de muestra, *n*, grandes.

La distribución normal estándar (cuando σ es conocido) o la T de Student (cuando σ es estimado por S calculada con los datos muestrales) pueden proveernos de la probabilidad de extraer aleatoriamente una media muestral que se posicione a un determinado número de desviaciones estándares de μ. Por ejemplo, las chances son de 1 en 20 de extraer una media que sea al menos 1.645 desviaciones estándares, más grande que la media poblacional si la distribución del estadístico es normal. Utilizando esta idea se construye el intervalo de confianza para la media poblacional a partir de la distribución muestral de $\bar{X}$ de la siguiente manera.

$$P(\bar{X} - Z_{1-\alpha/2}\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + Z_{1-\alpha/2}\sqrt{\sigma^2/n}) = 0.95$$

Luego los límites del intervalo de confianza para la media con nivel α=0.05 son:

$$LI = \bar{X} - 1.96\sqrt{\sigma^2/n} \quad y \quad LS = \bar{X} + 1.96\sqrt{\sigma^2/n}$$

En la práctica la varianza se estima desde la muestra, por lo que la estadística a usar debiera ser $T=\left(\frac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}\right)$ y no $Z=\left(\frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$. Los límites de los intervalos de confianza para la **Media** que reporta InfoStat son calculados como:

$$\text{LI} = \overline{X} - \text{T}_{1-\alpha/2}\sqrt{S^2/n} \quad \text{y} \quad \text{LS} = \overline{X} + \text{T}_{1-\alpha/2}\sqrt{S^2/n}$$

Podría suceder que no estemos seguros que las condiciones que garantizan la distribución de nuestro estadístico se cumplan y por ende que no conozcamos la distribución en el muestreo del estadístico que estamos usando para construir el intervalo de confianza. Para estas situaciones InfoStat permite seleccionar una técnica de construcción de intervalos no paramétrica basada en el procedimiento de remuestreo conocido como *bootstrap*. La técnica de *bootstrap* consiste en extraer al azar mediante un muestreo con reposición *B* muestras de tamaño *n* desde la muestra original de tamaño *n*. En cada una de las *B* muestras bootstrap (por defecto *B*=250) InfoStat calculará el estadístico de interés (en el ejemplo anterior, la media muestral) y ordenando ascendentemente las *B* estimaciones identifica los cuantiles que serán utilizados como límites del intervalo de confianza bootstrap del parámetro de interés. Así, seleccionando **Estimación por Bootstrap** los límites del intervalo bilateral con confianza (1-α)×100, corresponden a los percentiles (α/2)×100 y (1-α/2)×100 de la lista de estimaciones obtenidas en *B* muestras bootstrap extraídas de la muestra original.

Si se selecciona **Estimación paramétrica**, los intervalos son construidos bajo la teoría normal para los parámetros Media, Mediana y Varianza. Los límites de confianza son calculados a partir de las siguientes expresiones:

$$Media: \overline{X} \pm T_{1-\alpha/2}\sigma/\sqrt{n}$$

$$Mediana: Me \pm T_{1-\alpha/2}\frac{\pi}{2}\sigma/\sqrt{n}$$

$$Varianza: LI = \frac{S^2(n-1)}{\chi^2_{1-\alpha/2}};\ LS = \frac{S^2(n-1)}{\chi^2_{\alpha/2}}$$

donde $T_{1-\alpha/2}$ y $\chi^2_{1-\alpha/2}$ son los cuantiles (1-α/2) de la distribución T de Student y Chi cuadrado respectivamente.

En la construcción de intervalos de confianza para la proporción de éxitos, InfoStat utiliza directamente los cuantiles de la distribución Binomial (*n*,*P*) asociada al estadístico número de éxitos muestrales, con *n* igual al número de repeticiones y *P* la proporción de éxitos poblacional. Así, los intervalos de confianza son exactos. Varios textos de Estadística presentan los intervalos basados en la distribución asintótica normal de la proporción

muestral. Los mismos carecen de sentido si se pueden obtener los intervalos exactos y por ello no son ofrecidos por InfoStat.

Cuando se desea un intervalo para *P* (proporción de éxitos) y no se dispone de una variable binaria sino de una variable cuantitativa, InfoStat permite construir variables dicotómicas mediante la definición por parte del usuario del criterio utilizado para decidir si un valor de la variable cuantitativa debe ser considerado como éxito (1) o fracaso (0). Al seleccionar intervalo de confianza para la proporción se habilita una subventana: **Considera éxito valores**: >, >=, <, <= o = que un valor especificado por el usuario en el campo reservado para ingresar dicho valor.

Se puede optar por intervalos bilaterales o unilaterales tanto derechos como izquierdos. En el campo denominado **Confianza** el usuario debe ingresar el valor $(1-\alpha)\times100$, siendo α el riesgo de cometer error tipo I.

Ejemplo 3: Se dispone de los pesos de 20 individuos de una clase y se desea estimar por intervalo de confianza, con nivel de significación de 0.05, el peso promedio en la población de individuos de esa clase. El archivo Pesos contiene los 20 valores registrados en la muestra.

Activando el Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN UNA MUESTRA ⇒ INTERVALOS DE CONFIANZA, y seleccionando la variable *Pesos* como variable de análisis, al **Aceptar** aparece la ventana de opciones de **Intervalos de Confianza**. Si se selecciona **Media**, **Confianza 95%, Estimación paramétrica**, **Bilateral** se obtendrá el siguiente resultado:

Tabla 8: Resultados intervalo de confianza, estimación paramétrica. Archivo Pesos.

Intervalos de confianza
Bilateral
Estimación paramétrica

Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
Pesos	Media	68.50	0.80	20	66.82	70.18

Como se estableció una confianza de (1-0.05)100 = 95% con un tamaño muestral de n=20, el cuantil usado en la construcción del intervalo es $T_{1-\alpha/2}$ perteneciente a la distribución T de Student con (20-1)=19 grados de libertad (*g.l.*), este es $T_{1-\alpha/2}$ =2.09. Los límites son calculados como $68.50\pm2.09\times0.8$. Podemos concluir con un 95% de confianza que el intervalo [66.82;70.18] contiene el valor del peso promedio en la población desde la cual se extrajo la muestra.

Para el mismo ejemplo, si se solicita el intervalo de confianza *bootstrap*, seleccionando **Estimación por bootstrap**, se obtiene un resultado muy similar. Lo que sugiere que ésta debiera ser una técnica útil cuando no se conoce la distribución en el muestreo del estadístico utilizado.

Tabla 9: Resultados intervalo de confianza, estimación por bootstrap. Archivo Pesos.

Intervalos de confianza
Bilateral
Estimación por bootstrap (B=250)

Variable	Parámetro	Estimación	E.E.	n	LI(95%)	LS(95%)
Pesos	Media	68.47	0.82	20	66.78	69.95

Prueba de normalidad (Shapiro-Wilks modificado)

InfoStat permite probar si la variable en estudio tiene distribución normal. Las hipótesis de la prueba son:

H_0: las observaciones tienen distribución normal; versus H_1: las observaciones no tienen distribución normal

Para realizar esta prueba elegir Menú ⇒ ESTADÍSTICAS ⇒ INFERENCIA BASADA EN UNA MUESTRA ⇒ PRUEBA DE NORMALIDAD (SHAPIRO-WILKS MODIFICADO). La prueba se realiza con el estadístico de Shapiro-Wilks modificado por Mahibbur y Govindarajulu (1997).

Se presenta a continuación los resultados de la prueba de Shapiro-Wilks modificado para los datos de germinación del archivo *Atriplex*.

Tabla 10: Resultados prueba Shapiro-Wilks modificado. Archivo Atriplex.

Shapiro-Wilks modificado

Variable	n	Media	D.E.	W*	p (una cola)
Germinación	27	65.56	26.93	0.86	0.0033

En este caso hay evidencias para rechazar el supuesto de distribución normal (p<0.05).

Bondad de Ajuste (Kolmogorov)

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN UNA MUESTRA ⇒ BONDAD DE AJUSTE (KOLMOGOROV), permite probar si la muestra disponible se ajusta a un modelo distribucional teórico. Se supone que se dispone de una muestra aleatoria y que se desea probar si la distribución empírica se ajusta a alguna de las siguientes distribuciones: **Normal (m,v)**, **T de Student (v)**, **F de Snedecor (u,v)**, **Chi cuadrado (v)**, **Gamma (r,lambda)**, **Beta (a,b)**, **Weibull (a,b)**, **Exponencial (lambda)** o **Gumbel (a,b)**, (ver Capítulo Manejo de Datos). La distribución teórica debe ser completamente especificada (parámetros conocidos). Las hipótesis son:

$$H_0: G(x) = F_{teórica}(x) \text{ vs. } H_1: G(x) \neq F_{teórica}(x), \text{ para al menos una x}$$

donde G(x) es la función de distribución empírica (o de los valores observados) y $F_{teorica}(x)$ es la función de distribución teórica especificada por el usuario. La prueba es sensible a cualquier discrepancia entre las distribuciones (dispersión, posición, simetría, etc.). El estadístico se basa en la máxima diferencia entre las dos distribuciones y se define como la máxima distancia vertical entre G(x) y $F_{teorica}(x)$. El estadístico *D* de Kolmogorov es:

$$D = \sup_{x} \{|F_{teorica}(x) - G(x)|\}$$

InfoStat provee de valores *p* exactos para pruebas bilaterales obtenidos desde la distribución del estadístico *D*, la cual corresponde a una aproximación asintótica (Hollander y Wolfe, 1999).

La prueba de bondad de ajuste de Kolmogorov se usa cuando la función de distribución hipotetizada se especifica completamente, es decir que no se necesita estimar ningún parámetro desconocido desde la muestra. Por ello, seleccionada una distribución, InfoStat habilita tantos campos como parámetros la caracterizan para que el usuario introduzca el valor de cada parámetro. Cuando se introducen como valores de los parámetros las estimaciones muestrales, la prueba puede resultar demasiado conservadora. La prueba Chi cuadrado, disponible en el menú **Tablas de Frecuencias-Ajuste**, es menos afectada por las incorporaciones de estimaciones desde la muestra. Sin embargo, en algunas ocasiones la potencia de la prueba de bondad de ajuste Chi cuadrado es muy baja (Conover, 1999).

En la ventana **Resultados**, se especificará la variable y la distribución teórica seleccionada para el ajuste y las estimaciones por máxima-verosimilitud de los parámetros obtenidas a partir de las observaciones en la tabla de datos activa. Además se reporta el valor del estadístico (D) y el valor *p* correspondiente obtenido desde la aproximación asintótica de la distribución del estadístico.

Ejemplo 4: En un estudio sobre engorde de novillos se obtuvieron 150 registros de peso (kg). Se deseaba saber si la muestra se podía ajustar a un modelo distribucional normal con media 400 y varianza 400. El archivo Novillos contiene las observaciones del estudio. En la siguiente tabla se presentan los resultados de la prueba de bondad de ajuste de Kolmogorov.

Tabla 11: Resultados prueba Kolmogorov para bondad de ajuste. Archivo Novillos.

Prueba de Kolmogorov para bondad de ajuste

Variable	Ajuste	media	varianza	n	Estadistico D	p-valor
Peso	Normal(400,400)	399.57	478.23	150	0.08	0.3201

Valores de *p* menores al nivel de significación sugieren el rechazo de la H_0. Como el valor *p*=0.3201 se concluye que la muestra se ajusta al modelo distribucional propuesto.

Inferencia basada en dos muestras

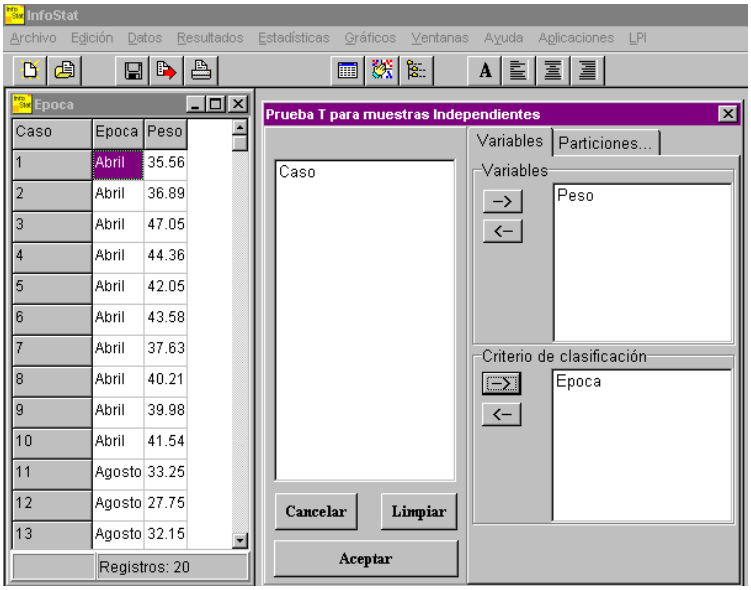
Prueba T para muestras independientes

Menú ⇒ ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA T (MUESTRAS INDEPENDIENTES), permite probar la hipótesis sobre la esperanza de la variable aleatoria definida como una diferencia de medias muestrales. Se asume que se dispone de dos muestras independientes, cada una desde una población o

distribución. La prueba puede ser vista como una herramienta para la comparación de medias (esperanzas) en dos poblaciones (distribuciones), es decir:

$$H_0: E(X_1) = E(X_2); \text{ versus } H_1: E(X_1) \neq E(X_2)$$

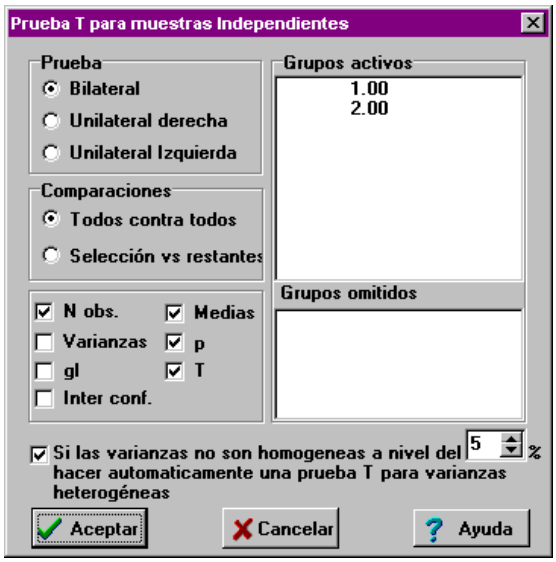
Al invocar esta prueba, en el selector de variables de la ventana **Prueba T para muestras independientes** se deberá especificar además de la variable respuesta en la subventana **Variables**, la variable que será usada para identificar ambas muestras en la subventana **Criterio de Clasificación**. La variable de clasificación debe permitir clasificar las observaciones en dos grupos, es decir identificar la procedencia (poblaciones) de cada muestra. Por ejemplo, en la siguiente ventana puede verse que en **Variables** se seleccionó el "Peso" y como **Criterio de clasificación** se eligió "Epoca".



Si se especifica más de una columna del archivo en la subventana **Criterio de clasificación** y/o más de una columna en la subventana **Variables**, InfoStat mostrará los resultados de las pruebas T asociadas a cada criterio de clasificación y/o a cada variable en estudio por separado. La ventana **Prueba T para muestras independientes** que se visualiza al **Aceptar** la selección de variables permite especificar el tipo de **Prueba**, el tipo de **Comparaciones,** la información a visualizar y requerir una prueba de homogeneidad de varianzas. La prueba puede ser bilateral, unilateral izquierda o unilateral derecha. Si existen más de dos grupos o muestras podrá obtener todas las pruebas T de a pares de medias seleccionando la opción **Todos contra todos**. InfoStat también permite comparar la media de uno de los grupos con respecto a la media de todos los grupos restantes activando la opción **Selección vs restantes**. Para indicar qué grupo fue seleccionado se deberá hacer doble *click* sobre el nombre del grupo en la subventana **Grupos activos**. Si se dispone de más de dos muestras y se desea omitir alguna se deberá hacer doble *click* sobre el nombre del grupo en la subventana grupos activos, teniendo activada la opción **Todos contra todos**, y dicho grupo

no será tenido en cuenta en el análisis por lo que se visualizará en la subventana **Grupos omitidos**. En cuanto a la información que se desea visualizar como resultado, el campo **Inter conf.** permite solicitar la construcción de un intervalo de confianza para la diferencia de medias poblacionales con coeficiente de confianza indicado por el usuario; los campos **T**, **gl** y **p**, cuando son activados, permiten la visualización del estadístico de la prueba, los grados de libertad de la distribución del mismo y el valor *p* de la prueba de hipótesis realizada.

Si se requiere la prueba de homogeneidad de varianzas InfoStat seleccionará el estadístico *T* para varianzas heterogéneas o para varianzas homogéneas según el resultado de la prueba. El usuario podrá especificar el **nivel de significación** a usar en la prueba de homogeneidad de varianzas.



En la ventana **Resultados** se muestran los valores de la variable de clasificación que determinan cada grupo y los tamaños de muestra en la población 1 y 2 (denotados por n_1 y n_2 respectivamente). Media(1), media(2), var(1) y var(2) son las medias y las varianzas de los grupos 1 y 2 respectivamente.

Ejemplo 5: En un estudio para analizar la evolución de tubérculos almacenados, se deseaba comparar dos épocas de cosecha: Abril y Agosto, las que determinan diferentes periodos de almacenamiento. La variable en estudio fue la pérdida de peso por deshidratación (en gr). El archivo Época contiene las observaciones del estudio. En la siguiente tabla se presentan los resultados.

Tabla 12: Resultados prueba T para igualdad de Medias. Archivo Epoca.

Prueba T para Muestras Independientes

Clasif	Var.	Grupo1	Grupo2	n(1)	n(2)	Media1	Media2	P(Var.Hom)	T	p	Prueba
Epoca	Peso	Abril	Agosto	10	10	40.89	26.65	0.6648	8.21	<0.0001	Bilat.

El valor $p<0.0001$ indica que hay diferencias entre las dos épocas en términos de las pérdidas de peso esperadas. Las medias muestrales sugieren una menor pérdida de peso promedio para el mes de Agosto.

Debido a que se solicitó la prueba de homogeneidad de varianzas, InfoStat contrasta las siguientes hipótesis: H_0: $\sigma_1^2=\sigma_2^2$ versus H_1: $\sigma_1^2\neq\sigma_2^2$

Para esta prueba se usa el estadístico F= $\frac{S_1^2}{S_2^2}$ que bajo H_0 se distribuye como una variable F con (n_1–1) y (n_2–1) grados de libertad. En la columna P(Var.Hom.) se presenta el valor *p* para la prueba de homogeneidad de varianzas. En este ejemplo, no se rechaza la hipótesis nula de homogeneidad de varianzas (nivel de significación nominal α=0.05).

Dado que la prueba de hipótesis indicó varianzas homogéneas, el estadístico *T*=8.21 es obtenido a partir de la siguiente expresión:

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

El valor *p* es calculado a partir de una distribución *T* de Student con (n_1+n_2-2) grados de libertad.

Cuando la hipótesis de homogeneidad de varianzas es rechazada, la prueba se basa en el estadístico:

$$T' = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{S_{\bar{X}_1 - \bar{X}_2}}$$

donde: $S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$.

En este último caso el valor *p* es calculado a partir de una distribución *T* de Student con ν grados de libertad, calculados a partir de la siguiente expresión:

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1+1} + \frac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$$

Prueba de Wilcoxon (Mann-Whitney U)

Menú ⇒ ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA DE WILCOXON (MANN-WHITNEY U), permite probar la hipótesis que dos muestras aleatorias independientes ($\{X_1,...,X_{n1}\}$ e $\{Y_1,...,Y_{n2}\}$), provienen de la misma población, usando el estadístico de Wilcoxon (Lehman, 1975). Esta prueba es equivalente a la prueba U de Mann Whitney para muestras independientes. Ambas son propuestas no paramétricas basadas en los rangos de las observaciones originales.

La hipótesis que se prueba es que las funciones de distribución subyacentes (F(x) y G(y)), tienen el mismo parámetro de posición. Bajo la hipótesis alternativa existe corrimiento (delta) de una distribución con respecto a la otra. Esto es:

$$H_0: F(x) = G(y)$$

y las posibles alternativas se basan en el modelo G(y) = F(x-δ), donde δ es el parámetro de corrimiento (bajo la hipótesis nula δ=0). La prueba está basada en el estadístico *W* el cual es la suma de los rangos en la muestra de menor tamaño. Los rangos son obtenidos a partir de la combinación de los datos de ambas muestras. El valor *W* en la tabla de resultados corresponde a una versión estandarizada del estadístico *W* basada en la distribución asintótica del mismo. Cuando el menor de los tamaños de las muestras bajo consideración es lo suficientemente grande, el estadístico *W* se distribuye como una variable normal estándar. InfoStat considera casos de empates modificando la varianza del estadístico (Hollander y Wolfe, 1999).

Para realizar esta prueba el archivo debe contener dos columnas, una indicando los valores de la variable y otra el criterio de clasificación como se mostró para la prueba T. Al **Aceptar** aparecerá la una pantalla donde el usuario debe seleccionar la información que desea contenga la ventana **Resultados**.

La ventana de resultados podrá presentar, para cada una de las muestras en consideración (grupo 1 y 2), el tamaño muestral (*n*(1) y *n*(2)), la media (media(1) y media(2)), la desviación estándar (DE(1) y DE(2)), la media de los rangos (R-media(1) y R-media(2)) y la mediana de la muestra original (mediana(1) y mediana(2)), el estadístico *W* y el valor *p*.

InfoStat permite obtener los valores *p* exactos (para ello, activar el campo **Exacta**), el valor p será calculado a partir de la distribución del estadístico para todas las muestras posibles. Cuando no se activa el campo **Exacta**, la versión estandarizada del estadístico *W* a partir de la cual se obtiene la prueba aproximada (si no hay empates) para muestras grandes es:

$$W^* = \frac{W - E(W)}{S(W)} \text{ donde: } E(W) = \frac{n(1)(n(2)+n(1)+1)}{2} \text{ y } S(W) = \sqrt{\frac{n(1)n(2)(n(2)+n(1)+1)}{12}}$$

Para el archivo *Época*, usando la prueba exacta se obtienen los siguientes resultados:

Tabla 13: Resultados prueba de Wilcoxon. Archivo Época.

Prueba de Wilcoxon (U Mann-Witney) para muestras independientes

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	DS(1)	DS(2)	W	p(2 colas)
Epoca	Peso	Abril	Agosto	10	10	40.89	26.65	3.58	4.15	155.00	0.0002

Dado que $p<0.05$, se rechaza la hipótesis nula, las distribuciones no tienen igual parámetro de posición.

Prueba de Wald- Wolfowitz

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ WALD WOLFOWITZ. Esta prueba no paramétrica puede aplicarse para determinar si dos muestras independientes han sido extraídas de la misma población contra la alternativa de que los dos grupos difieren en algún aspecto, ya sea tendencia central o variabilidad. El estadístico se basa en el número de rachas. Cuando los tamaños muestrales tienden a infinito, Wald y Wolfowitz demuestran que la estandarización del estadístico *R*, tiende a una distribución normal estándar (Lehmann, 1975). InfoStat usa la aproximación normal para la obtención de valores *p*.

Prueba de Van Der Waerden (puntuación normal)

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA DE VAN DER WAERDEN, permite comparar dos distribuciones de variables aleatorias continuas. Se asume que se dispone de dos muestras independientes. La hipótesis a probar es:

$$H_0: F(x) = G(y) \text{ versus } H_1: \delta \neq 0$$

donde δ es el parámetro de corrimiento.

Es una prueba competidora de la de Wilcoxon para probar igualdad de parámetros de posición en dos distribuciones.

El estadístico de Van Der Waerden es $c = \sum_{j=1}^{n} \phi^{-1}\left(\frac{S_j}{N+1}\right)$

donde *c*, es la suma desde $j=1,...,n$ de los cuantiles (S_j / N+1) de una distribución normal estándar acumulada (ϕ), *n* es el menor de los tamaños de las muestras en consideración y *N* es el total de observaciones.

Cuando el menor de los tamaños de las muestras bajo consideración es lo suficientemente grande, el estadístico *VW* (estandarización de *c*), se distribuye como una variable normal estándar (Hollander y Wolfe, 1999). InfoStat usa aproximación normal. La ventana de resultados presenta, para cada una de las muestras en consideración (grupo 1 y 2), el tamaño muestral (*n*(1) y *n*(2)), la media (media(1) y media(2)), la desviación estándar (DE(1) y DE(2)), la mediana de la muestra original (mediana(1) y mediana(2)), y los valores de *c* (suma Zscore(1) y suma Zscore(2)). Además, se puede solicitar el estadístico *VW* que se calcula como:

$$VW = \frac{c}{S(c)},$$

donde $S(c) = \sqrt{\frac{n(1)n(2)\sum_{i=1}^{N}\left\{\phi^{-1}\left(i/(n+1)\right)\right\}^2}{N(N-1)}}$ es el parámetro de corrimiento.

Prueba de Bell-Doksum (puntuación normal)

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA DE BELL-DOKSUM, permite comparar dos distribuciones de variables aleatorias continuas. Se asume que se dispone de dos muestras independientes. Es una prueba no paramétrica. La hipótesis a probar es:

$$H_0: F(x) = G(y) \text{ versus } H_1: \delta \neq 0$$

donde δ es el parámetro de corrimiento.

Para esta prueba se reemplazan los valores observados por puntuaciones normales (*Z*), se selecciona una muestra aleatoria de tamaño *n* de una tabla de valores para la distribución normal estándar. Los valores obtenidos se ordenan en forma ascendente y luego se reemplazan por los valores observados previamente ordenados respetando el orden. Esto se hace para cada grupo independientemente. InfoStat realiza sobre las puntuaciones normales (*Z*) una prueba T para dos muestras independientes, para muestras grandes usa aproximación normal. El estadístico con distribución aproximada normal estándar es calculado como:

$$Z = \frac{\bar{T}_1 - \bar{T}_2}{\sqrt{\frac{1}{n(1)} + \frac{1}{n(2)}}}$$

donde $\bar{T}_1$ y $\bar{T}_2$ representan los promedios de las puntuaciones normales en el grupo 1 y 2 respectivamente.

Ejemplo 6: Se realizó un experimento en el cual se registró el tiempo (en min.) de reacción ante una droga, para dos grupos de cobayos: Machos (1) y Hembras (2), (Archivo Bell-Doksum).

Tabla 14: Resultados de la prueba de Bell-Doksum. Archivo Bell-Doksum.

Bell-Doksum Normal Score

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	DS(1)	DS(2)	T	p(2 colas)
Grupo	Min.	1	2	5	5	26.40	44.00	5.68	7.62	-3.23	<0.0121

Prueba de Kolmogorov-Smirnov

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA DE KOLMOGOROV-SMIRNOV, permite comparar si dos muestras provienen de la misma distribución. Se supone que se dispone de dos muestras aleatorias independientes. Las hipótesis son:

H_0: F(x) = G(x) versus H_1: F(x) ≠ G(x), para al menos un x.

La prueba es sensible a cualquier discrepancia entre las distribuciones (dispersión, posición, simetría, etc.). El estadístico se basa en la máxima diferencia entre las dos distribuciones. El estadístico *KS* de Kolmogorov-Smirnov es:

$$KS = \frac{n(1)n(2)}{d} \max_{(-\infty<t<\infty)} \left\{ \left| F_{n(1)}(t) - G_{n(2)}(t) \right| \right\}$$

donde *n(1)* y *n(2)* son los tamaños muestrales y *d* es el máximo común divisor de *n(1)* y *n(2)*. InfoStat usa la aproximación normal basada en la distribución asintótica de *KS* adecuadamente normalizada (Hollander y Wolfe, 1999). Para cada muestra se puede solicitar además la media, la mediana y la desviación estándar (DE).

Prueba de Irwin-Fisher

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA DE IRWIN-FISHER, permite comparar dos muestras aleatorias que provienen de dos poblaciones independientes. Es un procedimiento para variables dicotómicas que se basa en la distribución hipergeométrica. Esta prueba permite contrastar la hipótesis de igualdad de proporciones de éxitos, p_1 y p_2, en ambas poblaciones:

H_0: $p_1=p_2=p_0$ versus H_1: $p_1 \neq p_2$

donde p_0 es un valor propuesto para el parámetro que se supone común a ambas distribuciones.

Ejemplo 7: Se está estudiando la respuesta de estudiantes a un método que mejora la gramática en el aprendizaje de un idioma. Un grupo de ellos fue asignado aleatoriamente al tratamiento experimental (A) y otro a un tratamiento control (B). Posteriormente, todos fueron sometidos a la misma evaluación que consistía en registrar el tiempo (en min.) que demoraban los alumnos en leer un párrafo y contestar la preguntas asociadas al mismo. Los resultados para los dos grupos en estudio conforman el archivo Idioma.

En la evaluación, cada estudiante se ubica por encima o por debajo de un nivel determinado, como por ejemplo la mediana de la muestra conjunta, que en este caso es 3.2.

El estadístico para la hipótesis de que la proporción de estudiantes por encima (o por debajo) del nivel es la misma para el tratamiento experimental y el control, es:

$$\hat{\Delta} = \hat{p}_1 - \hat{p}_2$$

con $\hat{p}_1 = \frac{t_1}{n(1)}$ y $\hat{p}_2 = \frac{t_2}{n(2)}$

donde $n(1)$ y $n(2)$ son los tamaños muestrales, t_1 y t_2 representan el número de elementos de $n(1)$ y de $n(2)$ que están por encima (o por debajo) del nivel. InfoStat calcula valores p exactos para esta prueba cuando el tamaño muestral no es superior a 50. Para tamaños muestrales grandes InfoStat usa la aproximación normal:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n(1)} + \frac{\hat{p}_0(1-\hat{p}_0)}{n(2)}}}$$

InfoStat permite construir automáticamente variables dicotómicas a partir de una variable cuantitativa seleccionada, mediante la definición por parte del usuario del criterio utilizado para decidir si un valor debe ser considerado como éxito (1) o fracaso (0). La comparación de cada valor con su media, mediana o un valor arbitrario o valores mayores, menores o iguales a ellos, puede ser indicada como criterio de dicotomización. El estadístico de la prueba será calculado como función de la variable dicotómica así creada.

Para el ejemplo presentado y eligiendo como opción **Valores mayores o iguales que la media** en la subventana **Mostrar la siguiente información**, se obtuvo la siguiente tabla:

Tabla 15: Prueba de Irwin-Fisher para comparación de proporciones. Archivo Idioma.

Prueba de Irwin-Fisher para comparación de proporciones

Criterio de corte: Valores mayores que la mediana(18.000000)

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	p1	p2	p1-p2	p(2 colas)
Grupo	VO	A	B	8	11	1.00	0.09	0.91	0.0001

En este ejemplo la proporción de casos por encima de la mediana es diferente para cada tratamiento (p=0.0001). La evidencia muestral indica mayor proporción para el Grupo A.

Prueba de la mediana

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA DE LA MEDIANA, permite obtener una prueba de homogeneidad de proporciones que es un caso especial de la prueba de Irwin-Fisher en el cual se utiliza la mediana como criterio para decidir si la observación debe ser considerada como un éxito o como un fracaso. Se supone que cada muestra se compone de observaciones independientes e idénticamente distribuidas que provienen de distribuciones continuas.

Para cada muestra se puede obtener P(X_1>Med) y P(X_2>Med), el número de observaciones que están por encima de la mediana calculada a partir de la muestra común (Med). Además, InfoStat provee las medias (media(1) y media(2)), las desviaciones estándar (DE(1) y DE(2)) y las medianas (mediana(1) y mediana(2)).

Las hipótesis son:

$$H_0: p_1=p_2=p_0 \text{ versus } H_1: p_1\neq p_2$$

donde p_0 es el valor hipotetizado para la proporción de casos por encima de la mediana de la muestra combinada, que se supone común para ambas muestras.

Para muestras grandes InfoStat usa la aproximación normal presentada para la prueba de Irwin-Fisher.

Siguiendo con el ejemplo del archivo *Idioma*, si se aplica la prueba de la mediana, se obtienen los siguientes resultados:

Tabla 16: Prueba de la mediana para dos muestras. Archivo Idioma.

Prueba de la mediana para dos muestras

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Med	P(X1>Med)	P(X2>Med)	p(2 colas)
Grupo	VO	A	B	8	11	18.00	1.00	0.09	<0.0001

Prueba para la diferencia de proporciones

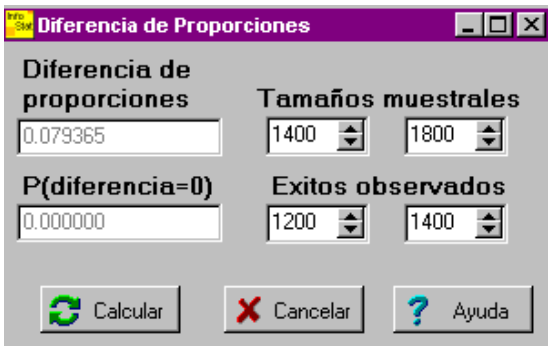
Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ DIFERENCIA DE PROPORCIONES, permite contrastar la hipótesis de igualdad de proporciones de éxito en dos poblacionales:

$$H_0: p_1=p_2 \text{ versus } H_1: p_1\neq p_2$$

Los valores *p* reportados se obtienen de la distribución exacta del estadístico de Fisher (Marascuilo, 1977).

Ejemplo 8: Una empresa encuestadora cree que la proporción de votantes (en la próxima elección), de la zona Norte es diferente a la proporción de votantes de la zona Sur. Se hace una encuesta entre los votantes de ambas zonas con los siguientes resultados:

Zona	Tamaño de la muestra	Número de personas que votarán en las elecciones
Norte	1400	1200
Sur	1800	1400



En este caso no se necesita tener un archivo de datos en InfoStat, simplemente se completan los campos de edición con la información requerida en cuanto a tamaños de las dos muestras disponibles y número de éxitos observados en cada uno de ellas. Al activar el botón **Calcular** se obtiene la diferencia de proporciones y el valor *p*, como puede verse en la pantalla.

Prueba T para observaciones apareadas

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA T (OBSERVACIONES APAREADAS). Permite probar la hipótesis de igualdad de medias cuando se toman observaciones de a pares desde las dos distribuciones que se comparan. Es decir que se dispone de una muestra de tamaño *n* de pares de observaciones, cada miembro de un par proveniente de una distribución. La prueba se basa en la distribución de la variable diferencia entre los pares de observaciones, *d*.

Si la hipótesis nula que se quiere probar es H_0: $\mu_1-\mu_2= 0$, esto implica $\mu_d = 0$, donde μ_d es la esperanza de la variable diferencia, para probar esta hipótesis el estadístico usado es:

$$T = \frac{\bar{d}}{S_d/\sqrt{n}} \sim T_{(n-1)}$$

donde: *n* es el número de pares, $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ y $S_d = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i-\bar{d}\right)^2}{n-1}}$ con d_i=diferencia entre las observaciones registradas en la i-ésima unidad muestral.

Observación: para poder realizar esta prueba InfoStat requiere de un archivo con dos columnas: una para las observaciones provenientes de la distribución 1 y otra para las de la distribución 2.

Ejemplo 9: Para estudiar el efecto de la polinización sobre el peso promedio de las semillas obtenidas, se efectuó un experimento sobre 10 plantas. La mitad de cada planta fue polinizada y la otra mitad no. Se pesaron las semillas de cada mitad por separado, registrándose de cada planta un par de observaciones. El archivo Poliniza contiene los valores registrados en el estudio, en la forma en que debieran ser ingresados en InfoStat.

Tabla 17: Resultados prueba T para observaciones apareadas. Archivo Poliniza.

Prueba T apareada

Obs(1)	Obs(2)	N	media(dif)	DE(dif)	T	p(2 colas)
Poliniz.	NoPoliniz.	10	0.45	0.17	8.42	<0.0001

En la salida, media(dif) y DE(dif) corresponden a la media y a la desviación estándar de la variable diferencia. El valor $p<0.0001$ sugiere el rechazo de la hipótesis H_0: $\mu_d =0$, es decir existen diferencias estadísticamente significativas entre los pesos de semillas que provienen de flores polinizadas y no polinizadas, la media de las diferencias de los pesos de flores polinizadas y no polinizadas es significativamente diferente de cero.

Prueba de Wilcoxon (para observaciones apareadas)

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA DE WILCOXON (OBSERVACIONES APAREADAS) permite obtener una prueba para la

comparación de dos distribuciones, que difieren eventualmente en su parámetro de posición, cuando se dispone de dos muestras con observaciones apareadas. Si **F**(.) y **G**(.) representan las funciones de distribución de X e Y respectivamente, la prueba de Wilcoxon prueba H_0: **F**(x)=**G**(y) versus H_1: **F**(x)=**G**(y-δ) con δ≠0, representando el parámetro de corrimiento. Es decir, el estadístico prueba la hipótesis de que las distribuciones de X e Y, siendo que en cada unidad muestral se registran valores de X y de Y (por ejemplo, reacción antes (X) y después (Y) de un tratamiento), son iguales excepto quizás por un cambio en el parámetro de posición. La prueba emplea la magnitud y signo de las diferencias entre los pares de observaciones. Dado un conjunto de observaciones pareadas (X_i,Y_i); i=1,...,n, el procedimiento calcula $D_i=(X_i-Y_i)$, los valores absolutos de las diferencias, y a ellos les aplica la transformación rango.

$$R_i=rango\,|X_i-Y_i|=\text{posición en la muestra ordenada de los}\,|D_i|$$

Posteriormente asocia a los rangos los signos de las diferencias originales. Esta prueba supone que la distribución de D_i es simétrica, que los D_i son mutuamente independientes con idéntica esperanza.

El estadístico de la prueba de Wilcoxon es la suma de los rangos correspondientes a $D_i>0$ y es denotado como T(+)=Suma R(+). InfoStat provee además la esperanza y la varianza de los rangos positivos bajo H_0. El valor *p* es obtenido por aproximación normal.

Ejemplo 10: Se desea comparar dos proyectos pedagógicos que se implementarán en dos escuelas (A y B), usando los resultados de una misma evaluación final realizada sobre un grupo de individuos que recibe enseñanza a través de estos dos métodos. Se eligen aleatoriamente de una población de interés (alumnos de cuarto año del secundario), 14 estudiantes, formándose 7 pares en base a sus notas promedio, las que se agrupan en 7 categorías. Los miembros de cada par fueron aleatoriamente asignados al método de enseñanza. Luego de un periodo de instrucción de un año se los evaluó. Los datos se presentan en el archivo Puntaje.

Tabla 18: Prueba de Wilcoxon. Archivo Puntaje.

Prueba de Wilcoxon (muestras apareadas)

Obs(1)	Obs(2)	N	Suma(R+)	E(R+)	Var(R+)	Bt p(2 colas)
Esc. A	Esc. B	7	27.00	14.00	34.63	0.0460

Para un nivel de significación α=0.05, existen diferencias estadísticamente significativas entre ambos métodos de enseñanza.

Prueba del signo

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA DEL SIGNO, permite obtener una prueba para la igualdad de esperanzas distribucionales en situaciones donde se dispone de dos muestras con observaciones apareadas. A diferencia de la prueba de Wilcoxon, trabaja sólo con el signo de las diferencias entre los *n* pares de observaciones (X,Y):

$$D_i = X_i - Y_i \quad \text{con } i=1,...,n$$

Si dicha diferencia es cero, se la considera empate y no se la incluye en el análisis. La hipótesis nula es: H_0: P (signos positivos)=P (signos negativos) = 1/2

Siguiendo con el ejemplo del archivo *Puntaje,* los D_i son {5,4,4,2,5,0,2}, cuya media es 3.14 y los signos correspondientes son: {+,+,+,+,+,0,+}. Los resultados de esta prueba para los datos del ejemplo, se presentan a continuación:

Tabla 19: Prueba del signo para muestras no independientes. Archivo Puntaje.

Prueba del signo

Obs(1)	Obs(2)	N	N(+)	N(-)	media(dif)	DE(dif)	p(2 colas)
Esc. A	Esc. B	7	6	0	3.14	1.86	0.0313

N(+) representa el número de diferencias positivas, N(-) el número de diferencias negativas, media(dif) y DE(dif) son respectivamente la media y la desviación estándar de la variable diferencia. Los resultados sugieren que existen diferencias entre ambos métodos de enseñanza (p=0.0313).

Observación: para poder realizar cualquiera de las pruebas que involucran observaciones apareadas, InfoStat requiere de un archivo con dos columnas, una para cada elemento del par de observaciones. El archivo tendrá tantas filas como pares de observaciones se hayan registrado.

Prueba F para igualdad de varianzas

Menú ESTADÍSTICAS ⇒ INFERENCIA BASADA EN DOS MUESTRAS ⇒ PRUEBA F DE IGUALDAD DE VARIANZAS permite contrastar la hipótesis de igualdad de varianzas de dos poblaciones.

InfoStat contrasta las hipótesis H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$ utilizando el estadístico $F = \frac{S_1^2}{S_2^2}$ que bajo H_0 se distribuye como una variable F con (n_1–1) y (n_2–1) grados de libertad.

Con los datos del Ejemplo 3 (archivo Época) se realizó una prueba F de homogeneidad de varianzas y los resultados se presentan en la siguiente tabla:

Tabla 20: Prueba F de homogeneidad de varianzas. Archivo Época.

Prueba F para igualdad de varianzas

Variable	Grupo(1)	Grupo(2)	n(1)	n(2)	Var(1)	Var(2)	F	p	prueba
Peso	{Abril}	{Agosto}	10	10	12.81	17.25	0.74	0.6648	Bilateral

El valor p=0.6648 indica que se acepta la hipótesis de homogeneidad de varianzas.

Análisis de la varianza

El Análisis de la Varianza (ANAVA), permite probar hipótesis referidas a los parámetros de posición (esperanza) de dos o más distribuciones. La hipótesis que se somete a prueba generalmente se establece con respecto a las medias de las poblaciones en estudio o de cada uno de los tratamientos evaluados en un experimento:

$$H_0: \mu_1 = \mu_2 = ... = \mu_a \qquad \text{con } i=1,...,a$$

donde *a*=número de poblaciones o tratamientos.

El ANAVA es un procedimiento que descompone la variabilidad total en la muestra (suma de cuadrados total de las observaciones) en componentes (sumas de cuadrados) asociados cada uno a una fuente de variación reconocida (Nelder, 1994; Searle, 1971, 1987).

En experimentos con fines comparativos, usualmente se realiza la aplicación de varios tratamientos a un conjunto de unidades experimentales para valorar y comparar las respuestas obtenidas bajo cada tratamiento. En este caso es deseable administrar eficientemente los recursos que permiten incrementar la precisión de las estimaciones de las respuestas promedio de tratamientos y las comparaciones entre ellas. Se entiende por *tratamientos* a la/s acciones que se aplican sobre las unidades experimentales y que son objeto de comparación. Los tratamientos pueden ser representados por los niveles de un factor o por la combinación de los niveles de dos o más factores (estructura factorial de tratamientos).

Uno de los principales objetivos en la planificación de una experiencia, siguiendo un diseño experimental, es la reducción del error o variabilidad entre unidades experimentales que reciben el mismo tratamiento, con el propósito de incrementar precisión y sensibilidad al momento de la inferencia, por ejemplo aquello relacionado a la comparación de efectos de tratamientos.

El diseño experimental es una estrategia de combinación de la estructura de tratamientos (factores de interés) con la estructura de unidades experimentales (parcelas, individuos, macetas, etc.), de manera tal que las alteraciones en las respuestas, al menos en algún subgrupo de unidades experimentales, puedan ser atribuidas solamente a la acción de los tratamientos excepto por variaciones aleatorias. Así, es posible contrastar (comparar) medias de tratamientos o combinaciones lineales de medias de tratamientos con el menor "ruido" posible.

Para realizar un ANAVA en InfoStat deben señalarse las variables del archivo que representan la o las *variables dependientes*, la o las *variables de clasificación* y la o las *covariables* en caso de que existan. La variable dependiente es la variable que se desea examinar (variable respuesta), por ejemplo rendimiento de un cultivo. Si más de una variable dependiente es especificada, InfoStat realizará el análisis de la varianza para cada una de las variables dependientes en forma separada.

Las variables de clasificación son las variables involucradas en el lado derecho de la ecuación del modelo estadístico del ANAVA y que representan factores o fuentes de variación que permiten separar o clasificar las observaciones del archivo en grupos. Normalmente existen factores que tienen que ver con la *estructura de tratamientos* del experimento y factores relacionados a la *estructura de las unidades experimentales*. Ambos deben ser señalizados como variables de clasificación.

Las variables indicadas como *covariables* (o variables concomitantes) representan variables aleatorias continuas cuyo valor varía con cada unidad experimental y que posiblemente están relacionadas linealmente con la variable respuesta. En situaciones donde se señala la presencia de una variable concomitante, InfoStat puede realizar *análisis de covarianza*, es decir, ajusta o remueve la variabilidad en la variable dependiente debida a la covariable antes de analizar las diferencias entre tratamientos.

A continuación se presenta la ventana **Análisis de la varianza** que permite elegir las variables indicadas para poder realizar este análisis.

En esta ventana el rendimiento representa la **Variable dependiente** y el cultivar la **Variable de clasificación**. Este ejemplo podría provenir de un experimento para comparar el rendimiento de dos o más cultivares y donde no se ha registrado ninguna covariable.

InfoStat usa el método de mínimos cuadrados para ajustar el modelo lineal general permitiendo especificar más de un criterio de clasificación y sus interacciones (factores cruzados) como así también estructuras de anidamiento (factores encajados). Con este tipo de modelos se pueden analizar experimentos con un solo factor o con múltiples factores o fuentes de variación (Cochran y Cox, 1957; Anderson y Mc Lean, 1974, Ostle, 1977; Hinkelmann y Kempthorne, 1994, Di Rienzo *et al.*, 2001).

Cuando se plantean las ecuaciones normales para obtener las estimaciones de los parámetros del modelo, se encuentran dependencias lineales por lo cual no existe una solución única para el sistema de ecuaciones (Graybill, 1961; Hocking, 1996). Para poder obtener una solución InfoStat usa las restricciones usuales: suma de los efectos de distintos niveles de un factor igual a cero. Esta clase de restricciones sobre los efectos de los factores en el modelo tiene una interpretación sencilla si se definen los efectos de los niveles de los factores como desvíos con respecto a la media general. Al imponer esas restricciones, se obtienen soluciones para los parámetros fijos del modelo.

Estas soluciones son utilizadas para obtener el valor *predicho* para cada observación. InfoStat calcula también los *residuos* para cada observación como la diferencia entre el valor observado y el predicho por el modelo. Los valores predichos y residuos de cada observación pueden ser anexados a la tabla de datos activa.

InfoStat trabaja, en este menú, con modelos de efectos fijos. Sin embargo, se permite la especificación de términos de error especiales para pruebas de hipótesis sobre términos del modelo. Así, el usuario puede tratar modelos de efectos fijos, aleatorios o mixtos si conoce las esperanzas de los cuadrados medios correspondiente a cada término del modelo.

Las sumas de cuadrados presentadas en las tablas de análisis de varianza son por defecto las *sumas de cuadrados de tipo III*. Estas sumas de cuadrados son llamadas *parciales* y reflejan la contribución de cada término del modelo, dado que todos los otros términos están también presentes en el modelo. Existe la opción de obtener las *sumas de cuadrados tipo I*. Las sumas de cuadrados de tipo I son llamadas *secuenciales*, y dependen del orden en que son declarados los términos del modelo ya que representan la reducción en la suma de cuadrados del error de cada término teniendo en cuenta la reducción debida a los términos ingresados anteriormente. Las sumas de cuadrados tipo I se usan cuando el orden de los términos del modelo se relaciona con una jerarquía que es útil para su interpretación. Por ejemplo, en modelos completamente encajados o en modelos con términos polinómicos.

Modelo

Los datos a analizar pueden provenir de ensayos realizados bajo diferentes diseños experimentales (completamente aleatorizado, bloques completos al azar, bloques incompletos balanceados, cuadrado latino, *crossover*, parcelas divididas, anidados etc.) (Snedecor, 1956; Ostle, 1977; Di Rienzo, *et al*., 2001). Las diferencias en el análisis de datos provenientes de distintos diseños son introducidas al especificar el modelo para la variable observada. InfoStat requiere la identificación del modelo lineal a utilizar en el análisis, en la ventana **Análisis de varianza,** solapa **Modelo**. Las variables declaradas como **Variables de clasificación** y como **Covariables** aparecerán automáticamente en la subventana **Especificación de los términos del modelo**, para que el usuario construya el modelo particular que desea ajustar. Los términos μ (media general) y ε (error aleatorio), presentes en todos los modelos de ANAVA, no necesitan ser especificados.

A continuación se presenta una breve descripción de las principales características de diseños experimentales corrientemente utilizados y la modalidad a seguir, en InfoStat, para la especificación del modelo asociado.

Diseño completamente aleatorizado

Se supone que las unidades experimentales son homogéneas, es decir no tienen estructura alguna. Los tratamientos se asignan completamente al azar a las unidades experimentales. El archivo de datos debe contener al menos dos columnas, una identificando al tratamiento (variable de clasificación) y otra a la variable respuesta (variable dependiente). El número

de repeticiones puede variar de un tratamiento a otro. El modelo lineal para la observación del tratamiento *i* en la parcela *j*, Y_{ij}, ajustado por InfoStat es:

$$Y_{ij} = \mu+\tau_i +\varepsilon_{ij}$$

donde:

Y_{ij} observación del tratamiento *i* en la parcela *j*

τ_i efecto del tratamiento *i*

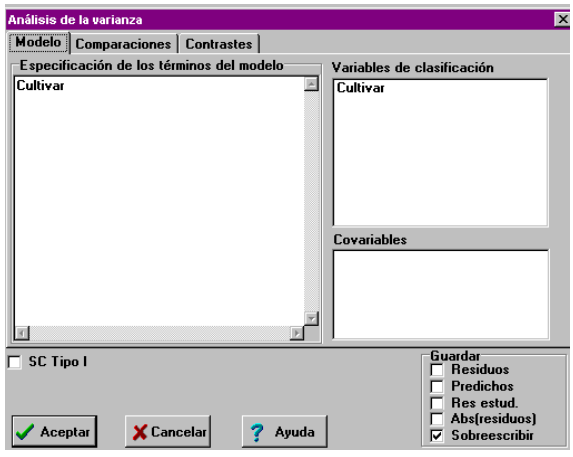
ε_{ij} término de error aleatorio asociado a la observación Y_{ij}

Usualmente se asume que el término de error se distribuye normalmente con media cero y varianza constante para toda observación.

Ejemplo 11: Para comparar 4 cultivares de maíz (tratamientos) se realiza un experimento bajo un diseño completamente aleatorizado con 10 repeticiones o parcelas por tratamiento. La variable de respuesta es el rendimiento. Los resultados se encuentran en el archivo Híbridos.

El archivo de datos *Híbridos* contiene dos columnas, una identificando al tratamiento (cultivares), y otra a la respuesta observada (rendimiento). Para realizar el análisis elegir Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA. Si en la ventana del selector de variables **Análisis de la varianza,** se declaró "cultivar" como **Variable de clasificación** y "rendimiento" como **Variable dependiente**, la siguiente ventana **Análisis de Varianza** señalará que la variable "cultivar" ha sido seleccionada como única variable de clasificación y esta variable también aparecerá en la subventana **Especificación de los términos del modelo** ya que se trata de un análisis a una vía de clasificación.

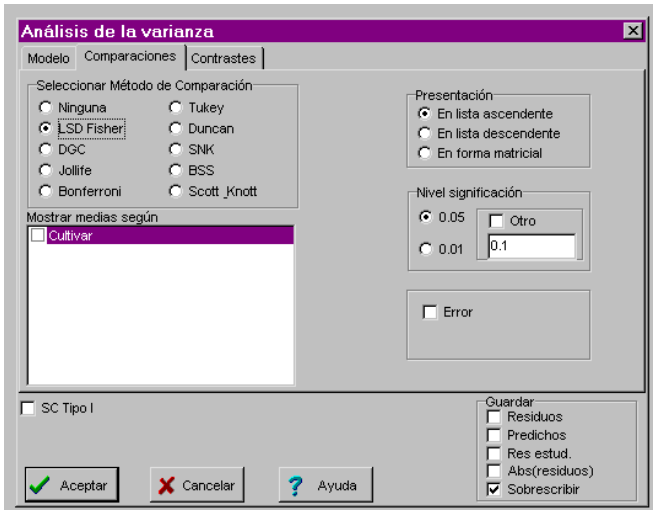
En esta ventana el usuario también podrá indicar si desea guardar los **residuos**, los valores **predichos** por el modelo, los **residuos estudentizados** y/o los **valores absolutos de los residuos** los cuales son útiles para la evaluación posterior del ajuste realizado con el modelo especificado (ver Supuestos). Cuando se solicitan estas medidas se crean automáticamente en el archivo de datos columnas conteniendo la información solicitada. El campo **Sobreescribir** debe ser activado cuando ya existen, en el archivo, columnas conteniendo residuos y valores predichos provenientes de una corrida anterior y no se desean conservar.



La ventana del análisis de la varianza presenta además de la solapa **Modelo**, las solapas **Comparaciones** y **Contrastes**. Estas permiten al usuario seleccionar el método de comparaciones múltiples de

medias que desea realizar a *posteriori* del análisis de varianza (ver Comparaciones múltiples) y establecer contrastes determinados *ad hoc* entre las medias de los distintos niveles de clasificación (ver Contrastes).

En la ventana que aparecerá al seleccionar la solapa **Comparaciones**, en **Medias a comparar** se elige el factor del cual se desean comparar las medias (en este ejemplo se desean comparar de a pares las medias de los cultivares, por tanto debe seleccionarse "Cultivar"). Por tratarse de un diseño a una vía de clasificación, cualquiera de las opciones disponible, "Cultivar" y **todas las medias**, realizan las comparaciones deseadas. Si existen varios factores, las opciones disponibles incluyen el nombre de cada factor y **todas las medias.** Seleccionado un factor en particular se obtendrán las comparaciones de a pares de las medias correspondientes a cada nivel del factor seleccionado. Seleccionando **todas las medias**, InfoStat reportara las comparaciones de medias para todos los tratamientos definidos por la combinación de los niveles de los factores intervinientes.



Como ejemplo, en este problema, se seleccionó el método de comparaciones múltiples propuesto por Fisher (**LSD Fisher)** para comparar las medias de los cultivares de a pares. En esta ventana se especificó también que los resultados sean presentados **En forma de lista**, en tal caso InfoStat reporta las medias ordenadas de menor a mayor y acompañadas por una letra, de manera tal que las medias que tienen la misma letra no muestran diferencias estadísticamente significativas entre ellas, al nivel de significación propuesto en el campo correspondiente (por defecto 0.05, pero se puede especificar 0.01 u **Otro**). La opción Otro no esta disponible para las pruebas DGC, Tukey, Duncan y SNK.

Tabla 21: Cuadro de análisis de la varianza para un diseño completamente aleatorizado. Archivo Híbridos.

Análisis de la varianza

Variable	N	R^2	R^2Aj	CV
Rend.	40	0.321	0.265	23.726

Cuadro de Análisis de la Varianza (SC Tipo III)

F.V.	SC	gl	CM	F	p
Modelo	10026.830	3	3342.277	5.677	0.003
Cultivar	10026.830	3	3342.277	5.677	0.003
Error	21194.845	36	588.746		
Total	31221.676	39			

```
Test: LSD Fisher Alfa: 0.05 DMS: 22.00731
Error: 588.7457 gl: 36
```

Cultivar	Medias	n		
2.00	76.68	10	A	
4.00	105.44	10		B
1.00	106.90	10		B
3.00	120.06	10		B

```
Letras distintas indican diferencias significativas(p<=0.05)
```

Para este ejemplo, el valor *p*=0.003 del ANAVA sugiere el rechazo de la hipótesis de igualdad de medias de tratamientos, es decir, existen diferencias estadísticamente significativas entre los cultivares considerando la variable rendimiento. De acuerdo a la prueba LSD de Fisher el cultivar 2 presenta diferencias estadísticamente significativas con respecto a los restantes, y por ser estos los de mayor rendimiento se recomienda cualquiera de ellos.

La verificación de las suposiciones realizadas sobre el término de error y la comparación de medias de tratamientos generalmente acompañan este tipo de salida (ver Supuestos de ANAVA, Comparaciones múltiples y Contrastes).

Diseño en bloques

Cuando existe variabilidad entre las unidades experimentales, grupos de unidades experimentales homogéneas pueden ser vistos como *bloques* para implementar la estrategia experimental conocida como *Diseño en Bloques*. El principio del bloqueo señala que las unidades experimentales dentro de cada bloque o grupo deben ser parecidas entre sí (homogeneidad dentro de bloque) y que los bloques debieran ser diferentes entre sí (heterogeneidad entre bloques). Es decir, el bloqueo o agrupamiento del material experimental debe ser tal que, las unidades experimentales dentro de un bloque sean tan homogéneas como sea posible y los bloques deben diseñarse para que las diferencias entre unidades experimentales sean explicadas, en mayor proporción, por las diferencias entre bloques. Cuando el diseño ha sido conducido en bloques, el modelo para cada observación debe incluir un término que represente el efecto del bloque al que pertenece la observación. Así, es posible eliminar de las comparaciones entre unidades que reciben el distinto tratamiento, variaciones debidas a la estructura presente entre parcelas (bloques).

Si cada bloque tiene tantas unidades experimentales como tratamientos y todos los tratamientos son asignados al azar dentro de cada bloque el diseño se denomina *Diseño en Bloques Completos al Azar* (DBCA). Se dice que el diseño es en *bloques completos* porque en cada bloque aparecen todos los tratamientos, y *al azar* porque dentro de cada bloque los tratamientos son asignados a las parcelas en forma aleatoria. Todas las parcelas de un mismo bloque tienen la misma probabilidad de recibir cualquiera de los tratamientos. La variación entre bloques no afecta a las diferencias entre medias, ya que cada tratamiento aparece el mismo número de veces en cada bloque. Este diseño permite mayor precisión que el completamente aleatorizado, cuando su uso está justificado por la estructura de las parcelas.

El siguiente modelo lineal puede ser postulado para explicar la variación de la respuesta, que en el bloque *j* recibe el tratamiento *i*, obtenida en un diseño en bloque con sólo un factor tratamiento:

$$Y_{ij} = \mu+\tau_i+\beta_j+\varepsilon_{ij} \qquad \text{con } i=1,...,a$$

donde μ corresponde a la media general, τ_i el efecto del i-ésimo tratamiento, β_j el efecto del j-ésimo bloque (j=1,...,b) y ε_{ij} es el error aleatorio asociado a la observación Y_{ij}. Comúnmente los términos de error se asumen normalmente distribuidos con esperanza cero y varianza común σ^2.

Otro supuesto que acompaña la especificación del modelo para un diseño en bloques se refiere a la aditividad (no interacción) de los efectos de bloques y de tratamientos.

Ejemplo 12: Se realizó un ensayo para evaluar el rendimiento en kg de materia seca por hectárea de una forrajera con distintos aportes de N_2 en forma de urea. Las dosis de urea probadas fueron 0 (control), 75, 150, 225 y 300 kg/ha. El ensayo se realizó en distintas zonas, en las que por razones edáficas y climáticas se podían prever rendimientos diferentes. Las zonas en este caso actuaron como bloques. El diseño a campo se ilustra en la Figura 1. Los datos se encuentran en el archivo Bloque.

Bloque I	225	300	75	0	150
Bloque II	300	150	75	0	225
Bloque III	75	0	300	225	150
Bloque IV	225	150	75	300	0

Figura 1: Asignación de tratamientos en un diseño en bloques completos aleatorizados. Archivo Bloque.

El archivo de datos para este análisis debe contener al menos tres columnas, una identificando al tratamiento (niveles de urea), otra a los bloques y otra a la respuesta observada (variable dependiente), en este caso el rendimiento. Para este análisis elegir Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA. Si en la ventana **Análisis de la varianza** se declara "tratamiento" y "bloque" como **Variables de clasificación** y "rendimiento" como **Variable dependiente,** la siguiente ventana es el selector de variables del **Análisis de Varianza** donde InfoStat señalará automáticamente que las variables "tratamiento" y "bloque" han sido seleccionadas como variables de clasificación y sus nombres aparecerán en la subventana **Especificación de los términos del modelo**.

Debido a que han sido declarados más de un término en el modelo, aparecerá automáticamente el botón **Agregar interacciones**. Para un diseño en bloques con sólo un factor tratamiento como el de este ejemplo, este botón no deberá activarse ya que el

supuesto de aditividad bloque-tratamiento justamente señala la falta de interacción entre los efectos de bloque y de tratamiento.

Al **Aceptar** se abrirá una ventana de **Resultados** conteniendo la siguiente información:

Tabla 22: Cuadro de análisis de la varianza para un diseño en bloques. Archivo Bloque.

Análisis de la varianza

Variable	N	R^2	R^2Aj	CV
Rendimiento	20	0.94	0.90	5.83

Cuadro de Análisis de la Varianza (SC Tipo III)

F.V.	SC	gl	CM	F	Valor p
Modelo	4494763.30	7	642109.04	24.88	<0.0001
Tratamiento	4291444.30	4	1072861.08	41.57	<0.0001
Bloque	203319.00	3	67773.00	2.63	0.0983
Error	309716.50	12	25809.71		
Total	4804479.80	19			

El valor $p<0.0001$ es menor al nivel de significación nominal de la prueba (suponga $\alpha=0.05$) y por tanto implica que el valor calculado del estadístico F a partir del experimento es mayor al valor teórico esperado bajo la hipótesis de igualdad de efectos de tratamientos (cuantil 0.95 de una distribución F con 4 y 12 grados de libertad). Luego se concluye con un nivel de significación del 0.05 que existen diferencias de rendimientos (kg de materia seca) bajo los distintos tratamientos o niveles de fertilización con urea.

La verificación de las suposiciones realizadas sobre el término de error y la comparación de medias de tratamientos generalmente acompañan este tipo de salida (ver Supuestos de ANAVA, Comparaciones múltiples y Contrastes).

En muchas situaciones no es posible asignar cada tratamiento en cada bloque. Cuando sólo un subconjunto de tratamientos está presente en cada bloque, el diseño se llama *Diseño en Bloques Incompletos*. InfoStat también permite ajustar modelos de ANAVA para Diseños en Bloques Incompletos. En este caso el diseño debe ser *balanceado* es decir, cada tratamiento debe estar presente, al menos en un bloque, con cada uno de los otros tratamientos. Esto es con el fin de proveer estimaciones de medias de tratamientos y de diferencias de medias de tratamientos con igual error estándar. El modelo que debe especificarse para un diseño de bloques incompletos es el mismo que para un DBCA. Al realizar el ANAVA, InfoStat ajusta la suma de cuadrados de los tratamientos por los bloques para remover el efecto de bloques desde las medias de tratamientos.

Diseño en cuadrado latino

En muchas situaciones, las unidades experimentales pueden ser agrupadas de acuerdo a más de un factor o fuente de variación independiente de los tratamientos. El diseño en cuadrado latino se utiliza para contemplar estructuras de parcelas donde intervienen dos factores de agrupamiento, comúnmente llamados factores *fila* y *columna* en el reconocimiento de fuentes de variación sistemática entre unidades experimentales. En un cuadrado latino cada tratamiento es aplicado una vez en cada fila y una vez en cada columna. Luego, si se

ensayan *a* tratamientos, un cuadrado latino se obtiene ordenando a^2 parcelas experimentales en un cuadrado de *a* filas y *a* columnas y asignando *a* parcelas a cada uno de los tratamientos de tal manera que en cada fila y en cada columna haya sólo una repetición de cada tratamiento como muestra la siguiente figura:

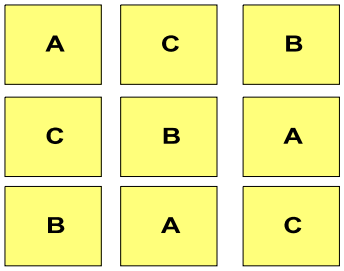


Figura 2: Diseño en cuadrado latino para un experimento en el que se ensayan tres tratamientos (A, B y C).

Este diseño impone un número fijo de repeticiones y cuando el número de tratamientos es grande, el experimento completo puede ser inmanejable. El número total de parcelas experimentales es igual al cuadrado del número de tratamientos.

El modelo lineal de ANAVA de un experimento con diseño en cuadrado latino es el siguiente:

$$Y_{ijk} = \mu + \tau_i + \delta_j + \gamma_k + \varepsilon_{ijk} \qquad \text{con } i, j, k = 1,\ldots,a$$

donde Y_{ijk} es la observación de la respuesta del i-ésimo tratamiento en la columna j-ésima y fila k-ésima, ε_{ijk} es el término de error aleatorio correspondiente a la observación del i-ésimo tratamiento en la columna j-ésima y fila k-ésima. En este modelo los parámetros δ_j y γ_k modelan los efectos de los factores asociados con variaciones en el sentido de las columnas y de las filas respectivamente. Los términos de error, usualmente, se asumen normalmente distribuidos con esperanza cero y varianza común σ^2. Otro supuesto que acompaña la especificación del modelo para un diseño en cuadrado latino se refiere a la aditividad (no interacción) de los efectos filas y columnas con los tratamientos.

Ejemplo 13: Se realizó un ensayo para evaluar el rendimiento en kg de materia seca por hectárea de una forrajera con distintos aportes de N_2 en forma de urea. Las dosis de urea probadas fueron 0 (control), 150 y 300 kg/ha. El ensayo se realizó en un potrero experimental con una cortina forestal al norte del mismo por lo que la luz recibida por las parcelas variaba de norte a sur. Además el lote presentaba una pendiente considerable de oeste a este. Se realizó un diseño en cuadrado latino, las variaciones de las parcelas en sentido norte-sur fueron consideradas como variaciones debidas al factor columna (luz) y aquellas en sentido oeste-este se asociaron al factor fila (pendiente). Los datos se encuentran en el archivo CuadLat.

El archivo de datos contiene cuatro columnas, una identificando al tratamiento (niveles de urea), otra al factor fila (pendiente), otra al factor columna (luz) y otra a la respuesta

observada (rendimiento). Para el análisis elegir Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA. Si en la ventana del selector de variables de **Análisis de la varianza** se declara "tratamiento", "fila" y "columna" como **Variables de clasificación** y "rendimiento" como **Variable dependiente,** la siguiente ventana **Análisis de Varianza** señalará que las variables "tratamiento", "fila" y "columna" han sido seleccionadas como variables de clasificación y aparecerán en la subventana **Especificación de los términos del modelo**. Al **Aceptar** se abrirá una ventana de **Resultados** conteniendo la siguiente información:

Tabla 23: Cuadro de análisis de la varianza para un diseño en cuadrado latino. Archivo CuadLat.

Análisis de la varianza

Variable	N	R^2	R^2Aj	CV
Rendimiento	9	0.998	0.992	1.364

Cuadro de Análisis de la Varianza (SC tipo I)

F.V.	SC	gl	CM	F	p-valor
Modelo	2698.00	6	449.67	161.88	0.0062
Fila	28.22	2	14.11	5.08	0.1645
Columna	754.89	2	377.44	135.88	0.0073
Tratamiento	1914.89	2	957.44	344.68	0.0029
Error	5.56	2	2.78		
Total	2703.56	8			

El valor p=0.029 menor al nivel de significación nominal (α=0.05) de la prueba para efecto de tratamientos, implica que el valor calculado del estadístico F a partir del experimento es mayor al valor teórico esperado bajo la hipótesis de igualdad de efectos de tratamientos (cuantil 0.95 de la F con 2 y 2 grados de libertad), luego se concluye con un nivel de significación del 0.05 que existen diferencias de rendimientos (kg. de materia seca producidos por la forrajera) bajo los distintos tratamientos o niveles de fertilización con urea.

Si los factores a controlar (estructuras de unidades experimentales) son tres, es decir, además de los efectos fila y columna existe otro efecto asociado a la estructura de parcelas, el diseño resultante suele denominarse *Greco-Latino*. Existen otras generalizaciones para este tipo de experimentos cuando el número de factores a controlar es superior a tres y cualquiera de los modelos asociados se pueden ajustar en InfoStat ya que el usuario puede adicionar tantos criterios de clasificación como sea necesario.

La verificación de las suposiciones realizadas sobre el término de error y la comparación de medias de tratamientos generalmente acompañan este tipo de salida (ver Supuestos de ANAVA, Comparaciones múltiples y Contrastes).

Diseños con estructura factorial de tratamientos

Estos diseños se utilizan para estudiar los efectos producidos por dos o más factores tratamiento y generalmente sus interacciones. En los ejemplos anteriores se definieron los tratamientos en relación a distintos niveles de un único factor de interés (estructura de tratamientos a una vía de clasificación). Cuando los tratamientos se definen mediante la

combinación de los niveles de dos o más factores de interés, se dice que el diseño experimental involucra una estructura factorial de tratamientos. La estructura factorial de tratamientos puede combinarse con distintos tipos de estructura de parcelas (completamente aleatorizado, bloques, etc.) para generar diversos diseños experimentales. InfoStat permite postular modelos que consideran estructura factorial de tratamientos ya sea en el marco de un diseño completamente aleatorizado, un diseño en bloques, un diseño en cuadrado latino, etc. Al especificar el modelo, los parámetros que hacen referencia a los efectos de tratamientos (que surgen de la combinación de dos o más factores) deben descomponerse en un conjunto de parámetros que dan cuenta de cada uno de los factores que intervienen para la definición de un tratamiento. Por ejemplo, si el tratamiento se define por la combinación de niveles del factor dosis y niveles del factor droga, será necesario especificar en el modelo los efectos dosis y tipo de droga por separado. Se pueden adicionar como términos del modelo todas las interacciones posibles entre los factores que intervienen. En cada caso habrá que juzgar si todas o sólo algunas de las interacciones adicionadas al modelo son requeridas. Si por ejemplo se realiza un ensayo con dos factores en un diseño en bloques completos al azar, no es necesario agregar las interacciones de bloque con cada uno de los dos factores de tratamiento.

InfoStat ajusta modelos para experimentos factoriales completos. En experimentos factoriales completos, se estudian todas las posibles combinaciones de los niveles de los factores (tratamientos) en cada repetición del experimento. Los modelos factoriales aditivos son aquellos en los que los términos que modelan la interacción están ausentes. Los modelos aditivos son usados para estudiar los efectos principales de los factores que intervienen en un proceso en el que se sabe que los factores no interactúan entre sí. El efecto principal de un factor se define como el cambio promedio en la respuesta producida entre cualquier par de niveles del factor considerado.

Para ejemplificar este caso se presenta un experimento factorial 2×2 (dos factores con dos niveles cada uno) en el que la interacción se supone ausente, el cual se ha dispuesto según un diseño completamente aleatorizado. Los factores se han designado como A y B y sus niveles como A_1,A_2 y B_1,B_2. Como existen 4 tratamientos (A_1B_1, A_1B_2, A_2B_1, A_2B_2) y suponiendo que estos no están repetidos, se tienen cuatro parcelas experimentales. Dado que el diseño es completamente aleatorizado la asignación de las parcelas a cada uno de los tratamientos es al azar. Un arreglo posible se presenta en la siguiente Figura:

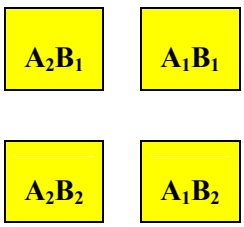


Figura 3: Experimento bifactorial sin repeticiones, bajo un diseño completamente aleatorizado.

El modelo para este experimento es el siguiente:

$$Y_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij}\ ; \qquad \text{con } i=1,2;\ j=1,2$$

donde Y_{ij} representa la respuesta al i-ésimo nivel del factor A y j-ésimo nivel de factor B, μ representa la media general, α_i el efecto que produce el i-ésimo nivel del factor A, β_j representa el efecto del j-ésimo nivel del factor B y ε_{ij} es el error aleatorio asociado a la observación ij-ésima. Los valores ε_{ij} usualmente se suponen normales, independientes, con esperanza cero y varianza común σ^2. Cuando los experimentos factoriales no tienen repeticiones, el analista debe suponer que los factores no interactúan para poder estimar la varianza del error experimental. Si este supuesto no se cumple entonces el experimento está deficientemente diseñado y las conclusiones del análisis pueden ser completamente erróneas, ya que la interacción será confundida con el error experimental.

Ejemplo 14: En un ensayo comparativo del efecto del estrés hídrico y salino sobre la germinación de Atriplex cordobensis, se sometieron lotes de semillas a cuatro niveles de potencial agua: 0, -0.5, -1.0 y -1.5 Mpa obtenidos mediante la aplicación al medio de dos osmolitos: polietilenglicol (PEG) y cloruro de sodio (ClNa). El experimento se condujo bajo un diseño completamente aleatorizado sin repeticiones. Los resultados se presentan en el archivo Factorial1.

El archivo de datos contiene tres columnas, una identificando al factor tratamiento A (estrés hídrico–potencial agua), otra al factor tratamiento B (estrés salino-osmolitos) y otra a la respuesta observada (germinación). Para este análisis elegir Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA. Si en la ventana del selector de variables del **Análisis de la varianza** se declara "Factor A" y "Factor B" como **Variable de clasificación** y "Germinación" como **Variable dependiente**, la siguiente ventana **Análisis de Varianza** señalará que las variables "Factor A" y "Factor B" han sido seleccionadas como variables de clasificación y aparecerán en la subventana **Especificación de los términos del modelo**. Como existe más de un factor de clasificación aparecerá automáticamente el botón **Agregar interacciones**. En este caso como no existen repeticiones la interacción no puede ser evaluada y por lo tanto no se debe activar este botón. Al **Aceptar** (sin agregar interacciones), se abrirá una ventana de **Resultados** conteniendo la siguiente información:

Tabla 24: Cuadro de análisis de la varianza para un experimento bifactorial. Archivo Factorial1.

Análisis de la varianza

Variable	N	R²	R²Aj	CV
Germinación	8	1.00	0.99	5.43

Cuadro de Análisis de la Varianza (SC Tipo III)

F.V.	SC	gl	CM	F	Valor p
Modelo	6568.50	4	1642.13	182.46	0.0007
Factor A	6518.50	3	2172.83	241.43	0.0004
Factor B	50.00	1	50.00	5.56	0.0997
Error	27.00	3	9.00		
Total	6595.50	7			

El valor $p<0.0004$, menor al nivel de significación nominal de la prueba ($\alpha=0.05$), para el efecto del factor A implica que, en el dominio estudiado, este factor tiene efecto estadísticamente distinto de cero sobre la germinación promedio. No sucede lo mismo para el factor B, ya que $p=0.0997$ es mayor al nivel de significación elegido.

La verificación de las suposiciones realizadas sobre el término de error y la comparación de medias de tratamientos generalmente acompañan este tipo de salida. Son necesarias pruebas de comparaciones múltiples de medias para el factor A, ya que el valor $p=0.0004$ solo rechaza la hipótesis de igualdad de medias entre los 4 niveles de potencial agua, pero no se conoce cual o cuales son diferentes (ver Supuestos de ANAVA, Comparaciones múltiples y Contrastes).

Puede encontrarse que la diferencia en la respuesta entre dos niveles de un factor no es la misma para distintos niveles de los otros factores, si esto ocurre se dice que hay una "interacción" entre los factores. Si el experimentador supone o sospecha que la respuesta a dos o más factores no se puede explicar como la suma de los efectos individuales de estos factores, entonces el modelo para el experimento factorial deberá incluir términos de interacción que den cuenta de este hecho. La inclusión de términos de interacción en el modelo conlleva la necesidad de tener repeticiones para cada tratamiento porque de otra forma no es posible estimar los parámetros adicionales (interacciones). Cuando el experimento tiene dos factores, existen sólo interacciones de primer orden, cuando tiene tres factores, existen interacciones de primer y de segundo orden y así sucesivamente existen estructuras factoriales que involucran interacciones de mayor orden.

El modelo para un experimento bifactorial con interacciones es una ampliación del modelo para el experimento bifactorial descripto anteriormente, excepto que incluye un conjunto adicional de parámetros, conocidos como parámetros de interacción:

$$Y_{ijk}=\mu+\alpha_i+\beta_j+\delta_{ij}+\varepsilon_{ijk} \qquad \text{con } i=1,2;\ j=1,2;\ k=1,..,n_{ij}$$

donde Y_{ijk} representa la respuesta de la k-ésima repetición en el i-ésimo nivel del factor A y j-ésimo nivel de factor B, μ representa una media general, α_i el efecto que produce el i-ésimo nivel del factor A, β_j corresponde al efecto del j-ésimo nivel del factor B, δ_{ij} el efecto adicional (interacción) para la combinación de los niveles *i* del factor A y *j* del factor B y ε_{ijk} es el error aleatorio asociado a la observación ijk-ésima.

Los términos ε_{ijk} que usualmente se suponen normal e independiente distribuidos con esperanza cero y varianza común σ^2. Debe notarse que el subíndice *k* se mueve entre 1 y n_{ij}, es decir, el número de repeticiones para el tratamiento puede ser distinto.

Ejemplo 15: En un estudio sobre la potencialidad forrajera de Atriplex cordobensis, un arbusto que crece en depresiones del chaco árido argentino, se evaluó la concentración de proteínas en hojas cosechadas en invierno y verano sobre plantas masculinas y femeninas. Para cada combinación de sexo y estación, se obtuvieron tres determinaciones del

contenido proteico medido como porcentaje del peso seco. Los resultados se encuentran en el archivo Factorial2.

El archivo de datos contiene tres columnas, una identificando al Factor A (sexo), otra al Factor B (estación) y otra a la respuesta observada (concentración de proteínas). Para este análisis elegir Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA. Si en la ventana del selector de variables del Análisis de varianza se declara "Factor A" y "Factor B" como Variable de clasificación y "Conc. Prot." como Variable dependiente, la siguiente ventana del selector de variables Análisis de Varianza señalará que las variables "Factor A" y "Factor B" han sido seleccionadas como Variables de clasificación y aparecerán en la subventana Especificación de los términos del modelo. Para incluir la interacción entre el Factor A y el Factor B, se deberá activar Agregar interacciones. Al Aceptar se abrirá una ventana de Resultados conteniendo la información provista en la siguiente tabla.

Nota: al activar **Agregar interacciones** se adicionarán automáticamente, como términos del modelo, todas las posibles interacciones entre las variables de clasificación. Para eliminar cualquier término que no se desea explicitar en la ecuación del modelo y que figura en dicha lista, seleccionar los mismos y presionar la tecla **Suprimir**. Para no agregar al modelo términos de interacción no deseados, en la subventana **Variables de clasificación** se pueden seleccionar aquellos términos cuyas interacciones se desean agregar al modelo y luego activar el botón **Agregar interacciones**.

Tabla 25: Cuadro de análisis de la varianza para un diseño bifactorial. Archivo Factorial2.

Análisis de la varianza

Variable	N	R^2	R^2Aj	CV
Conc.Prot.	12	0.93	0.91	6.30

Cuadro de Análisis de la Varianza (SC Tipo III)

F.V.	SC	gl	CM	F	Valor p
Modelo	198.00	3	66.00	37.71	<0.0001
Factor A	3.00	1	3.00	1.71	0.2268
Factor B	3.00	1	3.00	1.71	0.2268
Factor A*Factor B	192.00	1	192.00	109.71	<0.0001
Error	14.00	8	1.75		
Total	212.00	11			

Como puede observarse el valor *p* asociado a la interacción es altamente significativo, indicando que los factores estudiados no actúan independientemente. Por este motivo no se establecerán conclusiones sobre los efectos principales a partir de esta tabla ya que la presencia de interacción podría estar afectando las diferencias promedios. En este caso se deberán comparar las medias de los niveles del factor A dentro de los tratamientos que reciben el mismo nivel del factor B o viceversa (Comparar los niveles de B para cada nivel de A por separado).

Para obtener conclusiones acerca del efecto principal de un factor, en presencia de una interacción significativa, el experimentador debe examinar los niveles de dicho factor, manteniendo fijos los niveles de los otros. La siguiente figura muestra una gráfica, obtenida en InfoStat, de los valores medios en los cuatro tratamientos que permite interpretar fácilmente el resultado mostrado en el cuadro de análisis de la varianza.

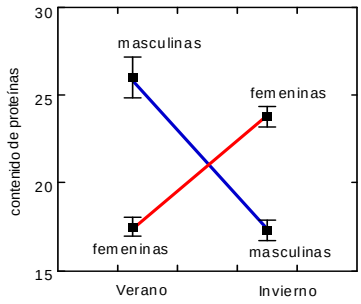


Figura 4: Media ± error estándar de la concentración de proteínas en hojas de Atriplex cordobensis para cada combinación de niveles de los factores sexo y época de cosecha.

Si alguna combinación de factores está ausente (celdas faltantes), InfoStat presentará automáticamente la suma de cuadrados de tipo I (secuenciales) e informará en la ventana de resultados que se trata de un diseño desbalanceado en celdas. Estas sumas son obtenidas para cada término del modelo secuencial si existe un subconjunto de datos que permite la estimación del contraste de interés (efectos principales e interacción) al menos para los datos existentes.

Diseño con estructura anidada de tratamientos

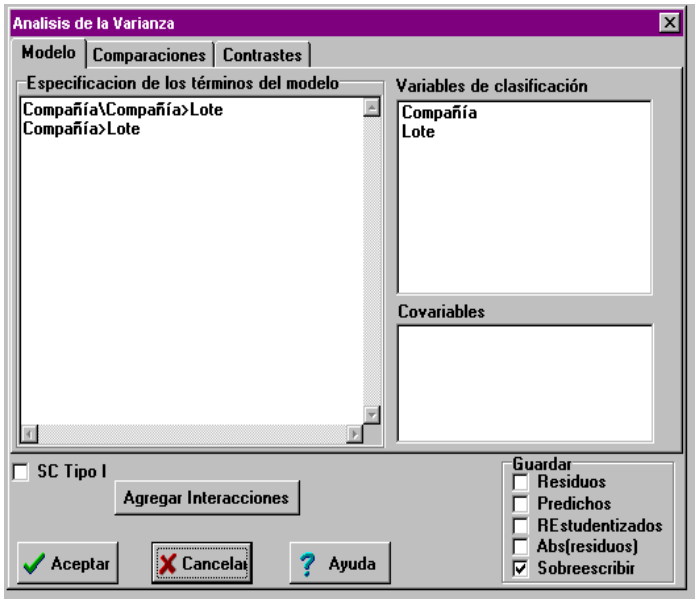
En algunos experimentos, con más de un factor tratamiento, los niveles de un factor (por ejemplo el Factor B), no representan lo mismo bajo diferentes niveles del otro factor (por ejemplo el Factor A) . Tal arreglo se conoce como *diseño anidado* y se dice que el Factor B está anidado en los niveles del Factor A si para cada nivel de A existe un conjunto particular de niveles de B. Si B está anidado en A, no tiene sentido estudiar la interacción entre A y B, ya que los niveles de B están siendo evaluados dentro de cada nivel de A. La tabla del ANAVA contempla la suma de cuadrados debida al Factor A, y la suma de cuadrados debida al Factor B dentro de A (InfoStat reconoce esta condición cuando se la declara como A>B). La suma de cuadrados del término A>B es la suma de las sumas de cuadrados del factor B y de la interacción B×A.

Ejemplo 16: Una compañía papelera compra papel por lotes a tres compañías madereras. La variable en estudio son los g/m^2 del papel obtenido en relación a la compañía y al lote. De cada compañía se examinan cuatro lotes de materia prima extraídos al azar y se hacen tres determinaciones por lote. Se puede decir que el factor "lote" está anidado dentro del factor "compañía", es decir, los lotes provenientes de distintas compañías no son los mismos. En este ejemplo, interesa conocer si hay diferencias entre compañías y si los lotes provenientes de las compañías son o no homogéneos en cuanto a la respuesta observada. Los datos se presentan en el archivo Anidado.

Para obtener la tabla de ANAVA usando InfoStat se deberá realizar el siguiente procedimiento: elegir menú ⇒ ESTADÍSTICAS, submenú ⇒ ANÁLISIS DE LA VARIANZA, y en la ventana del selector de variables del **Análisis de varianza** especificar

la **Variable dependiente** que en el ejemplo es "Valor", y las **Variables de clasificación** que son "Compañía" y "Lote". Al **Aceptar** se habilita la ventana de **Análisis de la Varianza** que contiene la solapa **Modelo** y aparecerán las dos variables de clasificación indicadas en la subventana **Especificación de los términos de modelo**. Para declarar los factores encajados o anidados se deberá escribir en esta subventana el nombre de los factores involucrados separados por el signo ">" (mayor), es decir en este caso "Compañía>Lote" que indica que el factor lote esta anidado en el factor compañía. El factor "Compañía", que está en el mayor nivel jerárquico, se declara por separado, así, en este ejemplo, en la ventana deben quedar los términos "Compañía" y "Compañía>Lote".

En este diseño, como en otros, para evaluar la significancia de las distintas fuentes de variación (factores del modelo), los términos de error a usar dependen de la naturaleza fija o aleatoria de los efectos del modelo. Por defecto, InfoStat calcula los estadísticos F sobre los factores declarados usando el término de error experimental, lo cual es válido si todos los efectos del modelo son fijos.



Si por ejemplo, el efecto compañía es fijo y el efecto lote aleatorio, el término de error para el factor "Compañía" es "Compañía>Lote". Para declarar esto en el modelo, se deberá indicar en la ventana **Especificación de los términos del modelo**, adicionando al factor Compañía el caracter "\" (barra invertida) y a continuación el término de error correspondiente, en este caso Compañía>Lote. La ventana deberá mostrar entonces los términos: Compañía\Compañía>Lote y Compañía>Lote.

Tabla 26: Cuadro de análisis de la varianza para un diseño anidado. Archivo Anidado.

Cuadro de Análisis de la Varianza (SC Tipo III)

Variable	N	R^2	R^2 Aj	CV
Valor	36	0.57	0.38	2.16

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor	(Error)
Modelo	84.97	11	7.72	2.93	0.0135	
Compañía	15.06	2	7.53	0.97	0.4158	(Compañía>Lote)
Compañía>Lote	69.92	9	7.77	2.94	0.0167	
Error	63.33	24	2.64			
Total	148.31	35				

En este ejemplo diremos que el factor compañía no tiene efecto significativo sobre los g/m^2 del papel (p=0.4158) y que la varianza entre los lotes dentro de al menos una compañía es distinta de cero (p=0.0167).

Diseño en parcelas divididas

Este tipo de diseño se usa frecuentemente en experimentos con más de un factor tratamiento y donde existen restricciones de aleatorización que impiden la asignación aleatoria de los tratamientos (combinación de factores) a las unidades experimentales. Son útiles cuando uno de los factores tratamiento necesita, para ser evaluado, parcelas o unidades experimentales grandes y el otro factor tratamiento se puede evaluar sobre unidades más pequeñas (subunidades).

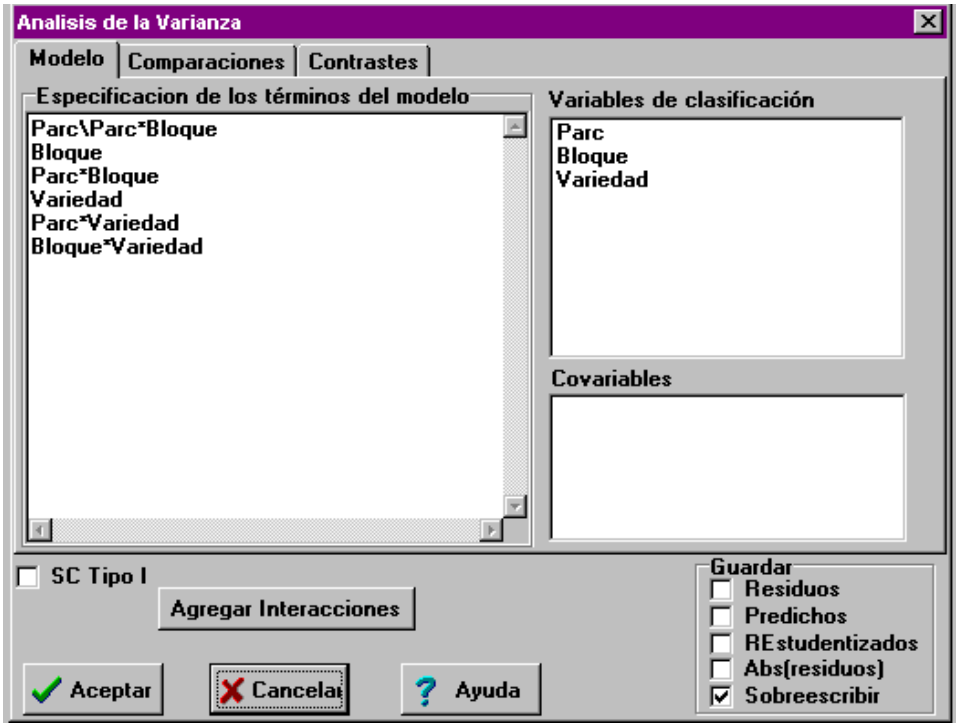
El diseño recibe el nombre de parcelas divididas ya que generalmente se asocia uno de los factores tratamiento (Factor A) a unidades experimentales de mayor tamaño (parcela principal) y dentro de cada nivel de este factor o lo que es lo mismo dentro de cada parcela principal se identifican "subparcelas" o parcelas de menor tamaño sobre las cuales se asigna al azar el segundo factor tratamiento (Factor B). A su vez podrían existir otros factores de bloqueo (no tratamiento) definiendo las estructuras de las parcelas. Así, entre otros, se puede tener un diseño en parcelas divididas con estructura de parcela completamente aleatoria o un diseño de parcelas divididas con estructura de parcelas en bloques, es decir, en cada uno de los *b* bloques del experimento se identifican parcelas principales y subparcelas. InfoStat permite manejar directamente cualquiera de estas estructuras experimentales.

Por ejemplo, considere un experimento para probar diferencias entre cuatro niveles de un factor A usando un diseño en bloques al azar y dos niveles de un segundo factor B que se asigna a subparcelas de la parcela principal asociadas con A. La aleatorización se realizará en dos etapas: primero se aleatorizan los niveles del factor A a las parcelas principales dentro de cada bloque, luego se aleatorizan los niveles del factor B en las subparcelas de cada parcela principal. Las restricciones a la aleatorización impuestas en este tipo de experimentos hacen necesario contar con más de un término de error. Si el diseño es en bloques completamente aleatorizados entonces la interacción bloque×parcela principal será el término de error para el efecto asignado a la parcela principal (error A). Si el diseño a nivel de las parcelas principales fuese completamente aleatorizado (las parcelas principales no se encuentran arregladas en bloques) se toma el efecto repeticiones dentro de parcela principal (parcela principal>repeticiones), como término de error para parcela principal. Los términos de error a utilizar dependerán de la estructura de parcelas y de la estructura de tratamientos. Si se tienen tres factores a evaluar cada uno con diferente tamaño de unidades experimentales entonces se tendrá un diseño en *parcelas sub-subdivididas* y por lo tanto se tendrán tres términos de error: error A (para parcela principal), error B (para subparcela) y el error experimental (para las sub-subparcelas).

Ejemplo 17: En un ensayo de trigo se dispusieron dos parcelas principales en tres bloques. Sobre las parcelas principales se aleatorizaron los niveles del factor riego y estas fueron divididas en cuatro subparcelas donde se aleatorizaron 4 variedades de trigo. La variable

en estudio fue el rendimiento medido en kg/parcela experimental. Para el factor "riego" (Factor A) se tienen dos niveles: secano (sin riego) y riego y para el factor "variedad" (Factor B) se usaron las siguientes variedades: Buck-Charrúa, Las Rosas-INTA, Pigue y Pro-INTA Puntal. Los datos (gentileza Ing. M. Cantarero, Facultad de Ciencias Agropecuarias, U.N.C.) se encuentran en el archivo ParcelaD.

Para realizar el análisis con InfoStat se debe proceder de la siguiente forma: elegir Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA, y en la ventana del selector de variables del **Análisis de varianza** especificar la **Variable dependiente** que en el ejemplo, es "Rendimiento" y **las Variables de clasificación** que son: "Parc" identificando la parcela principal (factor riego), "Bloque" y "Variedad". Al **Aceptar** se habilita la siguiente ventana de **Análisis de la Varianza**, allí en la solapa **Modelo** aparecen las variables de clasificación indicadas. Se deben agregar al modelo las interacciones Parc×Bloque (Error A, para evaluar el efecto parcelas) y la interacción Parc×Variedad ya que a pesar de las restricciones impuestas existe una estructura factorial de tratamientos.



La interacción entre variedades (subparcela) y bloques puede agregarse para probar el supuesto de no interacción. Algunos autores sugieren que luego de corroborada la ausencia de interacción bloque-subparcela, se puede presentar un análisis sin este término en el modelo para aumentar los grados de libertad del error. Con estas especificaciones se obtiene la siguiente salida:

Tabla 27: Cuadro de análisis de la varianza para un diseño en parcela dividida. Archivo ParcelaD.

Análisis de la Varianza

Variable	N	R^2	R^2 Aj	CV
Rendimiento	24	0.95	0.80	14.29

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor	(Error)
Modelo	389897.09	17	22935.12	6.46	0.0146	
Parc	276233.13	1	276233.13	55.24	0.0176	(Parc*Bloque)
Bloque	22912.97	2	11456.48	3.23	0.1117	
Parc*Bloque	10001.49	2	5000.74	1.41	0.3149	
Variedad	51095.57	3	17031.86	4.80	0.0491	
Parc*Variedad	18926.16	3	6308.72	1.78	0.2511	
Bloque*Variedad	10727.77	6	1787.96	0.50	0.7875	
Error	21286.97	6	3547.83			
Total	411184.06	23				

Como la interacción bloque×variedad no fue significativa (p=0.7875), la estructura de parcelas no interactúa con la estructura de tratamientos, se realizó un nuevo análisis sin ese término en el modelo.

Tabla 28: Cuadro de análisis de la varianza para un diseño en parcela dividida con parcelas principales repetidas en bloques completos aleatorizados. Archivo ParcelaD.

Análisis de la Varianza

Variable	N	R^2	R^2 Aj	CV
Rendimiento	24	0.92	0.85	12.39

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor	(Error)
Modelo	379169.31	11	34469.94	12.92	0.0001	
Parc	276233.13	1	276233.13	55.24	0.0176	(Parc*Bloque)
Bloque	22912.97	2	11456.48	4.29	0.0392	
Parc*Bloque	10001.49	2	5000.74	1.87	0.1957	
Variedad	51095.57	3	17031.86	6.38	0.0078	
Parc*Variedad	18926.16	3	6308.72	2.36	0.1223	
Error	32014.75	12	2667.90			
Total	411184.06	23				

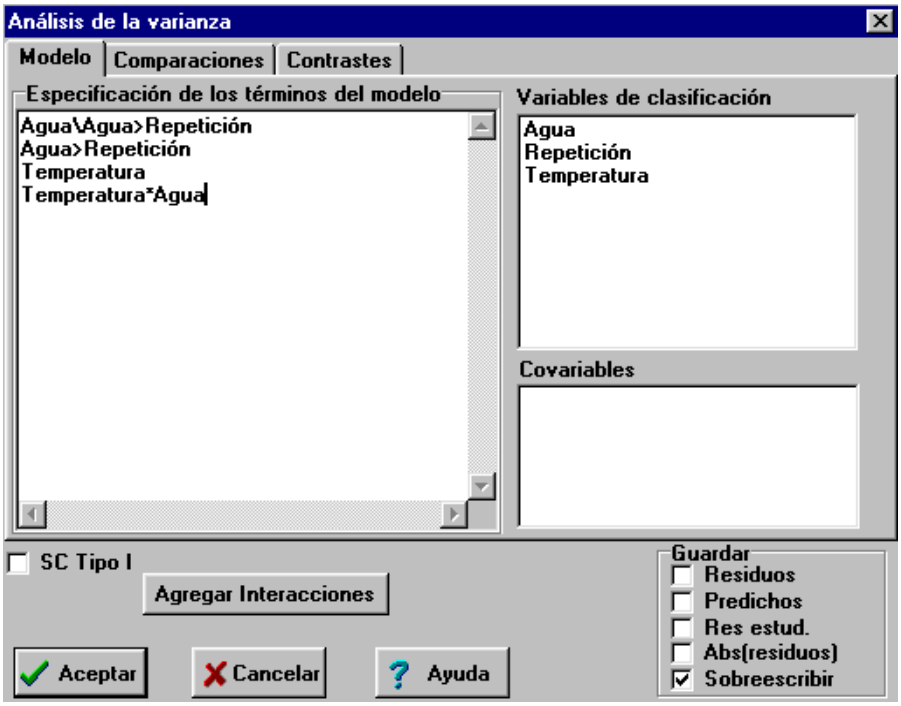
Los resultados sugieren que no hay interacción Parc×Variedad (p=0.1223), por lo que los resultados de los efectos principales pueden interpretarse directamente: existe efecto del riego (p=0.0176) y de la variedad (p=0.0078).

Para realizar las comparaciones y/o contrastes de medias entre los niveles de los factores intervinientes, InfoStat usará para cada término del modelo el error especificado en la columna (**Error)**.

Ejemplo 18: En un ensayo de resistencia de cartón se realizaron preparados de pasta básica con tres distintas cantidades de agua (50, 75 y 100 litros). Cada uno de los preparados (parcelas principales) se repitió tres veces en orden aleatorio a lo largo del tiempo. Luego, se dividieron los preparados en cuatro fracciones iguales (subparcelas) y se los sometió a distintas temperaturas de cocción (20, 25, 30 y 35 grados), las que fueron

asignadas al azar. La variable en estudio fue la resistencia del cartón obtenido. Los datos se encuentran en el archivo ParcelaDCA.

Para realizar el análisis con InfoStat se debe proceder de la siguiente forma: elegir Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA, y en la ventana del selector de variables del **Análisis de varianza** especificar la **Variable dependiente** que en el ejemplo, es "Resistencia" y **las Variables de clasificación** que son: "Agua" identificando la parcela principal (factor cantidad de agua), "Repetición" y "Temperatura". Al **Aceptar** se habilita la siguiente ventana de **Análisis de la Varianza**, allí en la solapa **Modelo** aparecen las variables de clasificación indicadas. Se deben agregar al modelo el término Agua>Repetición (Error A, para evaluar el efecto del agua) y la interacción Agua×Temperatura ya que a pesar de las restricciones impuestas existe una estructura factorial de tratamientos. A continuación se presenta la ventana con los términos del modelo propuestos para analizar este ejemplo.



Con estas especificaciones se obtiene la siguiente salida:

Tabla 29: Cuadro de análisis de la varianza para un diseño en parcela dividida con repeticiones completamente aleatorizadas para parcelas principales. Archivo ParcelaDCA.

Análisis de la varianza

Variable	N	R^2	R^2Aj	CV
resistencia	36	0.86	0.72	4.73

Cuadro de Análisis de la Varianza (SC Tipo III)

F.V.	SC	gl	CM	F	Valor p	Error
Modelo	933.84	17	54.93	6.27	0.0002	
Agua	522.75	2	261.37	25.50	0.0012	(Agua>Repetición)
Agua>Repetición	61.49	6	10.25	1.17	0.3649	
Temperatura	248.70	3	82.90	9.46	0.0006	
Agua*Temperatura	100.90	6	16.82	1.92	0.1324	
Error	157.66	18	8.76			
Total	1091.49	35				

Los resultados sugieren que no hay interacción Agua×Temperatura (p=0.1324), por lo que los resultados de los efectos principales pueden interpretarse directamente: existe efecto del agua (p=0.0012) y de la temperatura (p=0.0006).

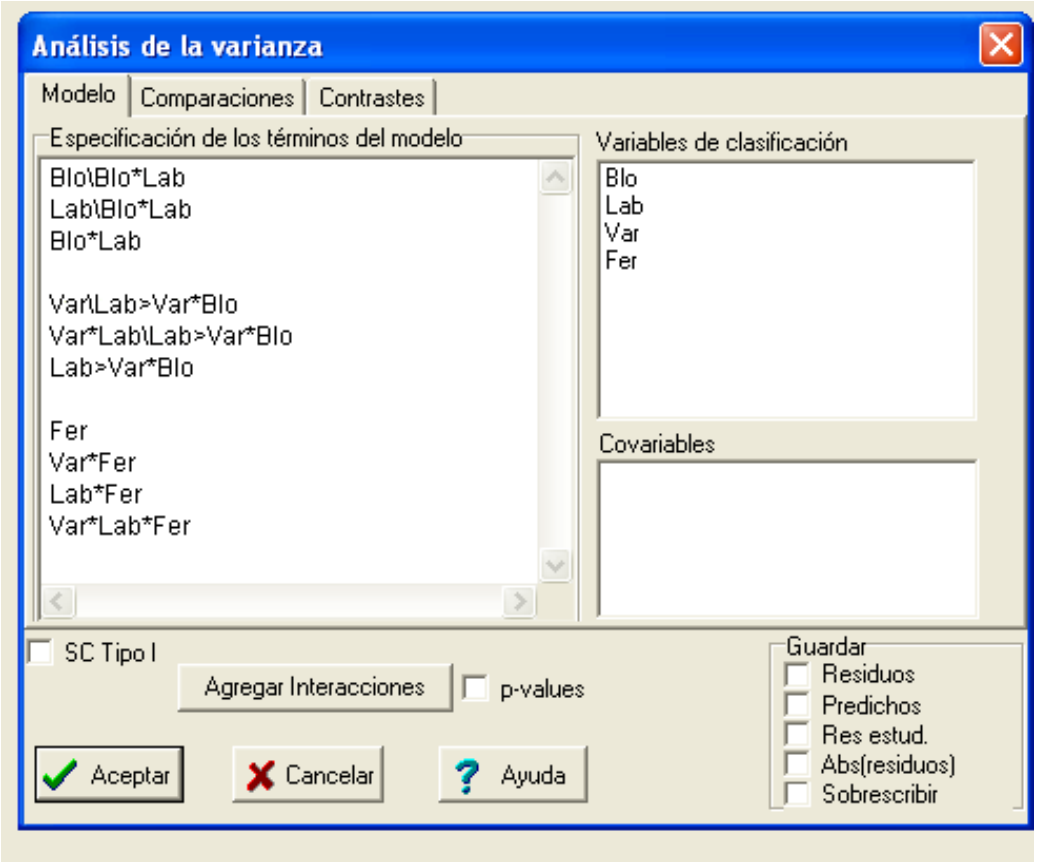
Diseño en Parcelas Subdivididas

Parcelas principales en BCA

Los datos en el archivo *parsubdiv.idb2* provienen de un diseño en bloques completos aleatorizados con 3 repeticiones (Blo). Cada bloque fue dividido en tres parcelas principales. En cada parcela principal (PP) se asignaron al azar tres métodos de labranza (Factor Lab, niveles Cero, Mínima y Convencional). Luego de la labranza, las parcelas principales fueron divididas en tres subparcelas (SP), y en cada una de ellas se asignaron al azar 3 variedades de maíz (Factor Var, niveles v1, v2 y v3). Por último, cada una de las subparcelas fue dividida en 4 sub-subparcelas (SSP), y en ellas se asignaron al azar 4 tipos de fertilizante (Factor Fer, niveles A, B, C, y D). La variable evaluada fue rendimiento de maíz (qq/ha). Para realizar el análisis se debe declarar como variable independiente a rendimiento y como variables de clasificación a Blo, Lab, Var y Fer.

El diseño en parcelas subdivididas implica tres instancias de aleatorización, por lo tanto para el análisis se deberán tener en cuenta tres errores diferentes: Uno para la parcela principal, uno para la subparcela y otro para las sub-subparcelas. En InfoStat solo se deben declarar los errores correspondientes a la PP y a la SP, ya que el tercero de los errores queda declarado por defecto. El error para la parcela principal es la interacción entre bloque y el factor que fue asignado en la PP, en este ejemplo, el método de labranza. El error para la SP esta dado por la interacción entre el bloque y el factor que esta en la subparcela, en este ejemplo Blo*Var, más la interacción triple de los factores de Bloque, PP y SP, en este caso Blo*Lab*Var. Esta suma puede reemplazarse en InfoStat por Lab>Var*Blo (Blo*Var + Blo*Lab*Var = Lab>Var*Blo)

Luego, en la solapa Modelo del menú Análisis de Varianza se deberá escribir lo siguiente:



En esta ventana se ha dejado un espacio entre los términos del modelo para PP, SP y SSP respectivamente para facilitar su visualización. En la solapa Comparaciones se solicito la prueba de Duncan. Al oprimir el botón Aceptar se obtendrá el siguiente resultado:

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
Rendimiento	108	0,97	0,95	2,63

Tabla 30: Cuadro de análisis de la varianza para un diseño en parcela sub dividida con parcelas principales repetidas en bloques completos. Archivo parsubdiv.idb2.

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor	(Error)
Modelo	1945,83	53	36,71	35,94	<0,0001	
Blo	636,72	2	318,36	216,25	0,0001	(Blo*Lab)
Lab	533,56	2	266,78	181,21	0,0001	(Blo*Lab)
Blo*Lab	5,89	4	1,47	1,44	0,2331	
Var	330,39	2	165,19	223,01	<0,0001	(Lab>Var*Blo)
Var*Lab	8,56	4	2,14	2,89	0,0690	(Lab>Var*Blo)
Lab>Var*Blo	8,89	12	0,74	0,73	0,7208	
Fer	400,78	3	133,59	130,77	<0,0001	
Var*Fer	3,39	6	0,56	0,55	0,7656	
Lab*Fer	7,56	6	1,26	1,23	0,3043	
Var*Lab*Fer	10,11	12	0,84	0,82	0,6247	
Error	55,17	54	1,02			
Total	2001,00	107				

Test:Duncan Alfa:=0,05

Error: 1,4722 gl: 4

Lab	Medias	n		
Convencional	36,33	36	A	
Mínima	37,61	36		B
Cero	41,56	36		C

Letras distintas indican diferencias significativas(p<= 0,05)

Test:Duncan Alfa:=0,05

Error: 0,7407 gl: 12

Var	Medias	n		
v3	36,75	36	A	
v1	37,86	36		B
v2	40,89	36		C

Letras distintas indican diferencias significativas(p<= 0,05)

Test:Duncan Alfa:=0,05

Error: 1,0216 gl: 54

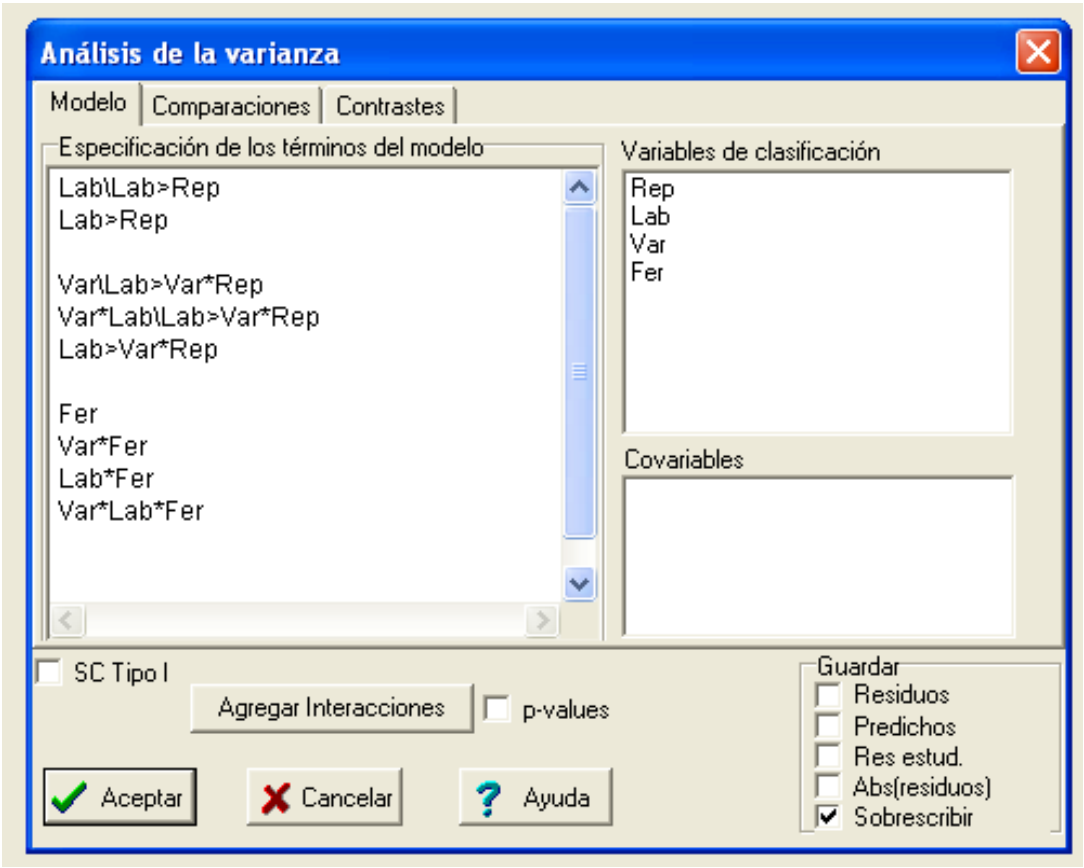
Fer	Medias	n		
A	36,52	27	A	
B	36,63	27	A	
C	40,41	27		B
D	40,44	27		B

Letras distintas indican diferencias significativas(p<= 0,05)

Este modelo contiene 3 factores de interés, por lo tanto lo primero que se debe evaluar es la interacción triple y las interacciones dobles. En este ejemplo ninguna interacción es significativa por lo tanto se pueden evaluar los efectos de los factores por separado. Existen diferencias entre método de labranza (p=0.0001) siendo la recomendada la labranza cero. También se encontraron diferencias entre variedades (p<0.0001) recomendándose la v2. Por último, hay diferencias entre los fertilizantes (p<0.0001), y se recomiendan indistintamente el C o el D.

Parcelas principales en diseño completamente aleatorizado

Si las parcelas principales son repetidas en un diseño completamente aleatorizado, lo único que cambia con respecto al modelo de bloques completamente aleatorizado es el error para la parcela principal (PP). En este caso, se deben declarar las repeticiones (Rep) y colocar en la solapa del Modelo los términos del factor de la PP y las repeticiones dentro de PP. Para el caso del ejemplo anterior (Archivo *parsubdiv.idb2*), suponiendo que las repeticiones de la PP no fueran en bloques sino completamente aleatorizadas, la ventana Modelo debería quedar como se muestra a continuación:



Comparaciones Múltiples

Usualmente, cuando los efectos de un factor en el ANAVA son considerados como no nulos, se implementa una prueba de comparaciones múltiples de medias. InfoStat provee las medias muestrales bajo cada una de las distribuciones que se comparan. El usuario puede indicar si desea comparar las medias de todos los tratamientos y en caso de diseños con estructura factorial de tratamientos, las medias de los niveles de cada uno de los factores que intervienen. Para analizar las diferencias de "a pares" entre las medias de las distribuciones que se comparan, es posible realizar una gran variedad de *pruebas a posteriori* o pruebas de comparación múltiple (Hsu, 1996; Hsu y Nelson, 1998).

La subventana **Comparaciones**, disponible en la ventana de ANAVA, permite solicitar procedimientos de comparaciones múltiples jerárquicos (basados en algoritmos de agrupamiento jerárquico) y tradicionales (Gonzalez, 2001). Los procedimientos tradicionales generalmente presentan una menor tasa de error tipo I que los procedimientos basados en conglomerados cuando se trabaja en experimentos que no tienen un buen control de los niveles de precisión usados para la comparación de medias. Con un número alto de medias de tratamiento, los procedimientos tradicionales pueden producir salidas de difícil interpretación ya que una misma media puede pertenecer a más de un grupo de medias (falta de transitividad). Por el contrario, los métodos jerárquicos para comparaciones de medias producen agrupamientos mutuamente excluyentes (partición del conjunto de medias de tratamientos). Di Rienzo *et al* (2001) y Gonzalez L.(2001) realizan la comparación simultánea de los métodos de comparaciones múltiples jerárquicos y no jerárquicos implementados en InfoStat.

Para cualquier procedimiento elegido, InfoStat permite definir el nivel de significación nominal usado para la prueba seleccionada. Además, se puede optar por el tipo de presentación de los resultados de las comparaciones múltiples (en forma de lista o en forma matricial). Si solicita presentación matricial, InfoStat presenta las comparaciones en una matriz cuya diagonal inferior tendrá como elementos las diferencias entre las medias y en la diagonal superior se presenta el símbolo "*" indicando los pares de medias que difieren al nivel de significación elegido. Si solicita presentación en lista, las comparaciones se muestran en una lista en la cual letras distintas indican diferencias significativas entre las medias que se comparan. El usuario puede también ingresar un valor correspondiente a la estimación del cuadrado medio de error (y sus grados de libertad) que desea sea utilizado en la comparación de medias. Cuando el casillero correspondiente ha sido activado, InfoStat no utiliza los términos de error usados en el último ANAVA como es habitual.

A continuación se presenta una breve descripción de los procedimientos disponibles en InfoStat.

Procedimientos tradicionales

Prueba LSD de Fisher

Compara las diferencias observadas entre cada par de promedios muestrales con el valor crítico correspondiente a la prueba *T* para dos muestras independientes. Cuando se trabaja con datos balanceados, esta prueba es equivalente a la prueba de la diferencia mínima significativa de Fisher, para toda comparación de medias de efectos principales. La prueba no ajusta el nivel de significación simultáneo, por lo cual la tasa de error por experimento puede ser mayor al nivel nominal, aumentando conforme aumenta el número de tratamientos a evaluar.

Prueba de Bonferroni

Se basa en una prueba *T* cuyo nivel ha sido ajustado de acuerdo a la desigualdad de Bonferroni. Permite controlar el error de tipo I de la inferencia simultanea de $c=k(k-1)/2$ contrastes de "a pares" basados, cada uno, en el estadístico *T* de Student, con *k*=número de tratamientos a comparar. El nivel de significación de cada contraste individual es ajustado de acuerdo al número de comparaciones realizadas. Cada contraste se realiza con un nivel de significación α/c.

Prueba de Tukey

Se basa en el estadístico de Tukey el cual calcula como valor crítico para la identificación de diferencias significativas, una cantidad (DMS) basada en el cuantil correspondiente de la distribución de rangos estudentizados. Cuando los tamaños de muestra son iguales, esta prueba controla la tasa de error por experimento, bajo hipótesis nulas completas o parciales. La prueba es más conservadora (error tipo I menor) que la prueba de Newman-Keuls o la de Duncan, en consecuencia puede perder potencia con respecto a ellas. Cuando los tamaños de muestras son desiguales, InfoStat implementa la modificación propuesta por Tukey-Cramer (Miller, 1981).

Prueba de Duncan

Conocida también como prueba de rangos múltiples, pertenece al tipo de pruebas conocidas como de etapas múltiples. Estas pruebas primero estudian la homogeneidad de todas las *k* medias a un nivel de significación α_k. Si se rechaza la hipótesis de homogeneidad para las *k* medias, se prueba homogeneidad en cada subconjunto de *k*-1 medias, usando un nivel de significación α_{k-1}, caso contrario el procedimiento se detiene. El procedimiento se repite hasta un nivel donde se encuentra que el subconjunto de medias involucradas es homogéneo. En términos generales el nivel de significación en la etapa i-ésima es: $\alpha_i=1-(1-\alpha)^{i-1}$. El método de Duncan controla que la tasa de error por comparación no supera el valor α nominal, sin embargo la tasa de error por experimento puede ser incrementada. Este incremento puede conducir a una disminución del error de tipo II, razón por la cuál algunos autores afirman que esta prueba es más potente.

Prueba de Student-Newman-Keuls (S.N.K)

Es también una prueba de rangos múltiples (ver Duncan). Las únicas diferencias con la prueba de Duncan es que el método SNK usa el mismo nivel de significación α en cada etapa. La prueba de Newman-Keuls controla la tasa de error por experimento, sólo cuando la hipótesis nula es completa, o sea cuando todas las medias son iguales. Sin embargo la tasa de error por experimento se aproxima a uno cuando el número de tratamientos se incrementa y existe mayor probabilidad de que la hipótesis nula sea verdadera sólo para un subgrupo de medias (hipótesis nulas parciales).

Procedimientos basados en conglomerados

Prueba de Di Rienzo, Guzmán y Casanoves (DGC)

Este procedimiento de comparación de medias (Di Rienzo, *et al.*, 2002) utiliza la técnica multivariada del análisis de conglomerados (encadenamiento promedio o UPGMA) sobre una matriz de distancia $D=\{d_{ij}\}=\left\|\bar{X}_i - \bar{X}_j\right\| / \sqrt{\frac{S^2}{n}}$ obtenida a partir de las medias muestrales.

Como consecuencia del análisis de conglomerado se obtiene un árbol binario en el cual puede observarse la secuencia jerárquica de formación de conglomerados. Si se designa como *Q* a la distancia entre el origen y el nodo raíz del árbol (aquel en el cual se unen todas las medias), InfoStat utiliza la distribución de *Q* bajo la hipótesis:

$$H_0: \mu_1 = \mu_2 = ... = \mu_k$$

para construir una prueba con nivel de significación α. Las medias (o grupos de medias) unidas en nodos que están por encima de *Q*, se pueden considerar estadísticamente diferentes para el nivel de significación α. El método presupone igual número de repeticiones por tratamiento, en caso contrario el algoritmo implementado utiliza la media armónica del número de repeticiones. Esta prueba controla bien la tasa de error tipo I por comparación manteniendo una potencia aceptable en experiencias bien conducidas (bajo CV para la diferencia de medias) y mejora su comportamiento general conforme aumenta el número de medias a comparar.

Prueba de Jollife

Esta prueba (Jollife, 1975) también consiste en dar un criterio de corte para un dendrograma que muestra las relaciones de similitud entre medias de tratamiento. La prueba consiste en aplicar un análisis de conglomerados basado en el algoritmo de encadenamiento simple a la matriz de valores *p* de la distribución de rangos estudentizados aplicada a las diferencias entre pares de medias. Aquellas medias que se unen más allá de 1-α se declaran estadísticamente diferentes. Esta prueba es la más conservadora dentro de los procedimientos jerárquicos.

Prueba de Scott y Knott

Scott y Knott (1974) fueron los pioneros en el uso de un criterio para la partición de conglomerados en el marco de una procedimiento de comparación de medias. La prueba implementada utiliza un método divisivo y el criterio de corte se basa en la distribución asintótica de un estadístico que recuerda una prueba F. Debido a que realiza sucesivas particiones con el mismo conjunto de datos el nivel de significación conjunto puede ser distinto del nominal por lo que debe ser tomado como un índice.

Prueba de Bautista et al (BSS)

Bautista *et al.* (1997) propusieron un algoritmo recursivo basado en la combinación de una técnica de conglomeración y un análisis de la varianza jerárquico. Tiene al igual que Scott y Knott el problema del nivel de significación de las comparaciones, que debe ser tomado como un índice, no obstante en estudios de simulación (Di Rienzo *et al.,* 2002) su comportamiento fue satisfactorio.

Contrastes

La subventana **Contrastes** disponible en la ventana principal del ANAVA permite obtener la significancia de contrastes postulados sobre los parámetros del modelo.

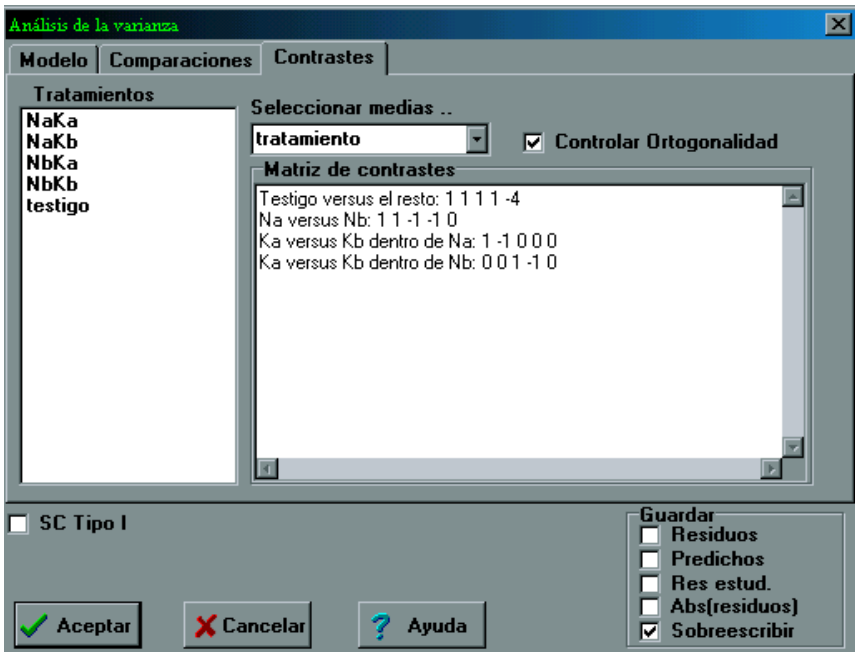
Un contraste es definido como una combinación lineal de los parámetros del modelo (Montgomery, 1991).

En el análisis de varianza los contrastes generalmente toman la forma $a_1M_1+ a_2M_2+...+ a_kM_k$ (donde los coeficientes a_i son constantes conocidas, al menos dos distintos de cero y su suma es cero, M_i es la i-ésima media poblacional.). Los contrastes permiten hacer comparaciones entre medias planeadas previamente al análisis de la varianza. Por ejemplo, si se tienen tres medias M_1, M_2 y M_3, el contraste 1 –1 0 comparará la M_1 con la M_2 y el contraste 2 -1 -1 es equivalente a comparar M_1 con la media de M_2 y M_3.

Si se desea plantear más de un contraste, para que las comparaciones sean independientes unas de otras, los contrastes deberán ser **ortogonales**. Dos contrastes son ortogonales si la suma de los productos de los coeficientes de ambos contrastes es cero. O sea, para $C_1 = a_1M_1 + a_2M_2+...+ a_kM_k$ y $C_2 = b_1M_1 + b_2M_2+...+ b_kM_k$, C_1 y C_2 son ortogonales si $a_1b_1+ a_2b_2+...+ a_kb_k = 0$. Tres o más contrastes son ortogonales si **todos los pares** de contrastes son **ortogonales**. InfoStat permite controlar la ortogonalidad de los contrastes propuestos marcando la opción **Controlar ortogonalidad**.

Ejemplo 19: En un ensayo para evaluar rendimiento en trigo se probaron cinco tratamientos: una combinación de N y K en dosis altas y bajas (NaKa, NaKb, NbKa, NbKb) más un tratamiento sin fertilizante (testigo). Obsérvese que los cinco tratamientos pueden verse como un arreglo factorial 2×2 con el agregado del testigo. Los datos se encuentran en el archivo Contraste1.

En la siguiente ventana se pueden ver un conjunto de contrastes ortogonales de interés. El primer contraste compara el testigo versus el promedio del resto de tratamientos. El segundo compara el promedio de los tratamientos con dosis baja de N con el promedio de los tratamientos de dosis alta. El tercer y cuarto contraste comparan los niveles del factor K en presencia de alta y baja cantidad de N respectivamente.



Para poder identificar mejor el contraste en la salida, InfoStat permite darle un nombre a cada contraste. Esto se logra si se coloca el nombre del contraste en la ventana **Matriz de contrastes** y a continuación, luego de poner ":" se escriben los contrastes de interés. Si el usuario no especifica un nombre para cada contraste, InfoStat los llamara Contraste 1, Contraste 2, y así sucesivamente, en función del orden en que fueron especificados.

En la siguiente tabla se muestra la salida correspondiente.

Tabla 31: Contrastes. Archivo Contraste1.

Análisis de la Varianza

Variable	N	R^2	R^2Aj	CV
rendimiento	15	0.945	0.923	4.191

Cuadro de Análisis de la Varianza (SC Tipo III)

F.V.	SC	gl	CM	F	p
Modelo	333.808	4	83.452	43.078	<0.0001
tratamiento	333.808	4	83.452	43.078	<0.0001
Error	19.372	10	1.937		
Total	353.180	14			

Contrastes

tratamiento	SC	gl	CM	F	p-valor

Testigo vs. el resto	186.92	1	186.92	96.49	<0.0001
Na vs. Nb	137.46	1	137.46	70.96	<0.0001
Ka vs.Kb dentro de Na..	9.41	1	9.41	4.86	0.0520
Ka vs.Kb dentro de Nb..	0.02	1	0.02	0.01	0.9300
Total	333.81	4	83.45	43.08	<0.0001

Coeficientes de los contrastes

tratamiento	Cont. 1	Cont. 2	Cont. 3	Cont. 4
NaKa	1.000	1.000	1.000	0.000
NaKb	1.000	1.000	-1.000	0.000
NbKa	1.000	-1.000	0.000	1.000
NbKb	1.000	-1.000	0.000	-1.000
testigo	-4.000	0.000	0.000	0.000

Como puede observarse, se rechazan las dos primeras hipótesis planteadas, es decir, el testigo difiere significativamente del resto (p<0.0001) y el nivel alto de N difiere del nivel bajo (p<0.0001). El contraste tres se insinúa como significativo (p=0.0520) y el cuatro no es significativo (p=0.9300) lo cual implica que no hay diferencias entre dosis baja y alta de K para niveles bajos de N.

A través de la subventana **Contrastes**, es posible especificar también pruebas para tendencias polinómicas cuando niveles del factor tratamiento son cuantitativos y equidistantes. Más que comparar niveles de a pares sobre todos los niveles del factor, es más informativo en estos casos investigar si existen tendencias en la respuesta debidas al incremento de los niveles del factor tratamiento. La tendencia podría verse reflejada por un incremento (disminución) lineal, cuadrático, cúbico etc. Los coeficientes de los contrastes a utilizar deben corresponderse con los coeficientes de los polinomios que modelan el tipo de tendencia que se hipotetiza. Con *a* niveles del factor tratamiento a analizar, es posible postular *a*-1 polinomios ortogonales de orden 1,...,*a*-1. Coeficientes para polinomios ortogonales para varios números de tratamientos pueden ser encontrados en Montgomery (1991).

Por ejemplo, si se tienen 5 niveles equidistantes para un factor, la suma de cuadrados de contrastes para los efectos lineales, cuadráticos, cúbicos y de orden cuatro, descomponen la suma de cuadrados de tratamiento, en cuatro contrastes con un grado de libertad para cada uno. Los coeficientes de los contrastes, en este ejemplo serían:

Tabla 32: Coeficientes para contrastes ortogonales.

Lineal	Cuadrático	Cúbico	Orden Cuatro
-2	2	-1	1
-1	-1	2	-4
0	-2	0	6
1	-1	-2	-4
2	2	1	1

Ejemplo 20: Para estudiar el efecto de una nueva formulación química para el control de un insecto se realizó una experiencia en la que se probaron la nueva formulación (nueva) y la formulación estándar (estándar) con tres dosis de producto activo cada una (15, 20 y 25 mg/l). La variable respuesta es el número promedio de insectos muertos por planta. Los datos se encuentran en el archivo Contraste2.

A continuación se presenta la tabla de ANAVA para este ejemplo:

Tabla 33: Contrastes. Archivo Contraste2.

Análisis de la varianza

Variable	N	R^2	R^2 Aj	CV
respuesta	12	0.88	0.78	22.35

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	160.25	5	32.05	8.99	0.0093
droga	33.81	1	33.81	9.49	0.0217
dosis	29.14	2	14.57	4.09	0.0758
droga*dosis	97.30	2	48.65	13.65	0.0058
Error	21.38	6	3.56		
Total	181.63	11			

Como puede observarse hay interacción dosis con droga (p=0.0058), por lo cual no se puede hacer inferencias sobre los efectos principales droga y dosis. Por este motivo se plantearon contrastes abriendo el término de interacción, es decir, estudiando un factor dentro de los niveles del otro. Debido a que es de interés conocer las tendencias lineales y cuadráticas de las dosis para cada una de las formulaciones, se realizaron contrastes polinómicos ortogonales para dosis dentro de cada formulación. Los contrastes y sus coeficientes se presentan a continuación:

Tabla 34: Contrastes. Archivo Contraste2.

Contrastes

droga*dosis	SC	gl	CM	F	p-valor
Contraste1	63.85	1	63.85	17.92	0.0055
Contraste2	1.62	1	1.62	0.45	0.5257
Contraste3	9.78	1	9.78	2.74	0.1487
Contraste4	51.20	1	51.20	14.37	0.0091
Contraste5	33.81	1	33.81	9.49	0.0217
Total	160.25	5	32.05	8.99	0.0093

Coeficientes de los contrastes

droga*dosis	Cont. 1	Cont. 2	Cont. 3	Cont. 4	Cont. 5
estandar:15.00	-1.00	1.00	0.00	0.00	1.00
estandar:20.00	0.00	-2.00	0.00	0.00	1.00
estandar:25.00	1.00	1.00	0.00	0.00	1.00
nueva:15.00	0.00	0.00	-1.00	1.00	-1.00
nueva:20.00	0.00	0.00	0.00	-2.00	-1.00
nueva:25.00	0.00	0.00	1.00	1.00	-1.00

El contraste 1 prueba tendencia lineal dentro de la formulación estándar y resultó significativo (p=0.0055), mientras que el contraste 2 prueba tendencia cuadrática y no resultó significativo (p=0.5257). El contraste 3 prueba tendencia lineal para las dosis de la nueva formulación y no resultó significativo (p=0.1487). El contraste 4 para la tendencia cuadrática en la nueva formulación no resulta significativo (p=0.0091). El contraste 5 prueba si hay diferencias entre la nueva formulación y la estándar y resulta significativo (p=0.0217). Este último resultado coincide con el resultado obtenido mediante el ANAVA

para el factor droga. La diferencia en las tendencias polinómicas de la respuesta a través de las dosis dentro de cada droga explica la interacción droga×dosis detectada en la tabla original.

Supuestos del ANAVA

El análisis de varianza es sensible a las propiedades estadísticas de los términos de error aleatorio del modelo lineal. Los supuestos tradicionales del ANAVA implican errores independientes, normalmente distribuidos y con varianzas homogéneas para todas las observaciones. Además, para diseños involucrando estructuras de parcelas en bloque se supone que existe aditividad bloque-tratamiento, es decir, los bloques tienen un efecto aditivo sobre todos los tratamientos y no interactúan con estos.

La verificación de los supuestos subyacentes se realiza en la práctica a través de los predictores de los términos de error aleatorio que son los residuos aleatorios asociados a cada observación.

InfoStat permite obtener, en el marco del ANAVA, los valores de **residuos**, **predichos**, **residuos estudentizados** y el **valor absoluto de los residuos**. Al seleccionar una o más de estas opciones se añadirán al archivo de datos columnas conteniendo los valores de cada una de las opciones elegidas. El residuo asociado a la observación ij-ésima (simbolizado como e_{ij}) es la diferencia entre el valor observado y el valor predicho por el modelo para la respuesta en la unidad experimental ij-ésima. A partir de los residuos y sus transformaciones se puede verificar el cumplimiento de los supuestos de normalidad, homogeneidad de varianzas y aditividad bloque-tratamiento mediante pruebas gráficas y/o formales (pruebas de adecuación del modelo). InfoStat permite obtener residuos en la escala de la variable observada (e_{ij}=diferencia entre valores observados y valores predichos por el modelo). Al activar el campo **Residuos**, InfoStat creará una columna denominada RDUO_"nombre de la variable" en la tabla de datos activa. También se puede activar el campo **Res estud**. para obtener los residuos estudentizados definidos como:

$$RE = \frac{e_i}{S\sqrt{1-h_{ii}}}$$

donde S^2=cuadrado medio del error y h_{ii} es el "leverage" de la i-ésima observación (Hocking, 1996). En este caso InfoStat creará una columna llamada RE_ "nombre de la variable" en la tabla de datos.

El casillero **Sobrescribir** se puede activar en situaciones donde se procesa más de un ANAVA sobre el mismo conjunto de datos y se desea guardar sólo el último conjunto de valores solicitados. Si este casillero no está activado, InfoStat generará tantas columnas nuevas como análisis de varianza se soliciten, numerando consecutivamente cada nueva variable agregada a la tabla.

Por lo general, en la práctica, los supuestos del ANAVA no se cumplen con exactitud. En caso de que haya evidencia de faltas graves de cumplimiento de los supuestos, el modelo y/o la estrategia de análisis podría no ser adecuado.

A continuación se presentan algunas estrategias que pueden conducirse en InfoStat para probar si se cumplen las suposiciones del ANAVA. En la ejemplificación de las pruebas de supuestos, se usaron los datos del archivo *Bloque*.

Normalidad: seleccionando los residuos como variable de análisis, una de las técnicas más usadas es construir un Q-Q plot normal. Mediante esta técnica se obtiene un diagrama de dispersión de los residuos obtenidos versus los cuantiles teóricos de una distribución normal. Si los residuos son normales y no hay otros defectos del modelo, se alinearán sobre una recta a 45°.

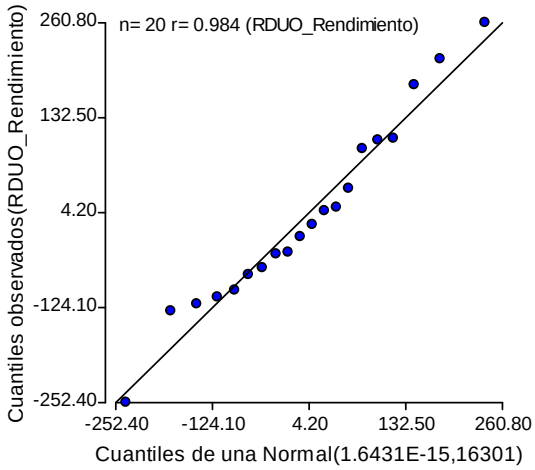


Figura 5: Q-Q plot (normal) obtenido a partir de un modelo con errores normales. Archivo Bloque.

Habiendo corrido un ANAVA y guardando los residuos, se debe seleccionar del Menú GRÁFICOS, de la barra de herramientas de InfoStat, opción Q-Q plot (normal) y usar como variable a los residuos del modelo. En la figura anterior se agregó, mediante la opción **Mostrar recta Y=X**, la recta de ajuste para estos residuos.

InfoStat permite realizar una prueba de hipótesis sobre normalidad, en el Menú ⇒ ESTADISTICAS ⇒ INFERENCIA BASADA EN UNA MUESTRA ⇒ PRUEBA DE NORMALIDAD (SHAPIRO-WILKS MODIFICADO). Allí, seleccionando los residuos como variable de análisis se obtiene el estadístico *W** de Shapiro-Wilks modificado por Mahibbur y Govindarajulu (1997).

Tabla 35: Prueba de normalidad. Archivo Bloque.

Shapiro-Wilks (modificado)

Variable	n	Media	D.E.	W*	p (una cola)
RDUO_Rendimiento	20	0.00	127.67	0.96	0.7824

Las hipótesis que se someten a prueba son:

H_0: los residuos tienen distribución normal versus H_1: los residuos no tienen distribución normal.

En este caso no hay evidencias para rechazar el supuesto de distribución normal (p=0.7824).

Homogeneidad de varianzas: Cuando los errores son homocedásticos, haciendo un gráfico de dispersión de residuos versus valores predichos se debe observar una nube de puntos sin patrón alguno (patrón aleatorio). Si el gráfico muestra estructura habrá indicios para sospechar sobre el cumplimiento del supuesto. Un patrón típico que indica falta de homogeneidad en las varianzas, se muestra en la Figura 6. Para el ejemplo que se trabaja en esta sección, el gráfico de residuos versus predichos se muestra en la Figura 7, no se observa tendencia que indique falta de cumplimiento del supuesto de homogeneidad de varianzas.

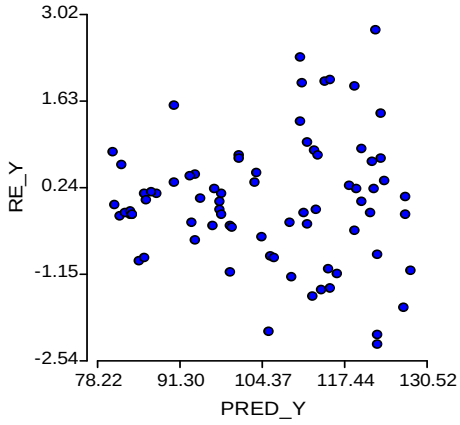


Figura 6: Gráfico de residuos en función de predichos en un ejemplo con falta de homogeneidad de varianzas.

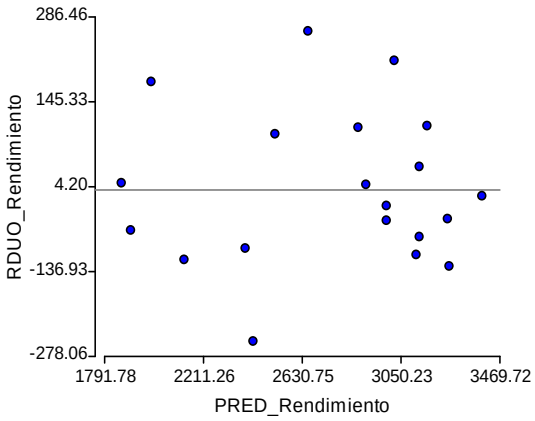


Figura 7: Gráfico de residuos en función de predichos. Archivo Bloque.

Otra estrategia para la validación del supuesto de homocedasticidad para el factor tratamientos, es la prueba de Levene (Montgomery, 1991). Si bien esta prueba fue desarrollada para diseños completamente aleatorizados (a una vía de clasificación), se puede extender su uso a modelos más complejos. La prueba consiste en realizar un análisis de la varianza usando como variable dependiente el valor absoluto de los residuos. Este análisis se debe realizar con un modelo a una vía de clasificación.

Las hipótesis que se someten a prueba son:

H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_a^2$ versus H_1: Al menos dos varianzas son distintas

donde σ_i^2 es la varianza del tratamiento *i*, *i*=1,...,*a*.

Si el valor *p* del factor tratamiento de este ANAVA es menor al valor de significación nominal se rechaza la hipótesis de varianzas homogéneas, caso contrario el supuesto de igualdad de varianzas puede ser sostenido. InfoStat no tiene implementada esta prueba como tal en la sección de las pruebas de hipótesis, pero se puede construir fácilmente ya que se pueden guardar automáticamente los valores absolutos de los residuos con la opción **abs(residuos)**.

Independencia: Para verificar el supuesto de errores independientes, se puede realizar un gráfico de dispersión de los residuos en función de la variable que se presume puede generar dependencias sobre las observaciones. Un ejemplo clásico es la secuencia en el tiempo en que los operarios realizaron las observaciones; si la técnica de medición y/o observación puede ser afectada por la fatiga de operario, los residuos puede no ser independientes de la secuencia de toma de datos. La estructura de dependencia puede relacionarse a la forma en que se recolectaron los datos. Una tendencia a tener agrupados residuos positivos y/o negativos indica la presencia de correlación o falta de independencia. En general, un buen proceso de aleatorización asegura el cumplimiento del supuesto de independencia.

Aditividad bloque-tratamiento: Además de los supuestos clásicos del ANAVA a una vía de clasificación (sin estructura de parcelas), es decir errores independientes e idénticamente distribuidos (i.i.d.) normales con media cero y varianzas homogéneas, en los diseños con estructura de parcela, se supone que ésta no interactúa con la estructura de tratamientos, es decir, sus efectos deben ser aditivos. No debe existir interacción entre los componentes de la estructura de diseño y los componentes de la estructura de tratamientos, ya que se asume que la relación existente entre tratamientos es consistente de bloque a bloque (excepto por variaciones aleatorias). Otra forma de entender este supuesto es pensar que los bloques o grupos de parcelas homogéneas no tienen influencia sobre las diferencias entre los tratamientos. En caso que no sea así debería utilizarse un diseño experimental que considere la interacción entre los factores usados para el bloqueo y los tratamientos. Para explorar este supuesto puede ser de utilidad graficar los valores de la variable (o los residuos) en el eje Y y los tratamientos en el eje X y utilizar conectores entre los puntos del gráfico que provengan de un mismo bloque (particionar por bloque). La existencia de cruzamientos o falta de paralelismo de los perfiles graficados sugiere falta de aditividad o presencia de interacción bloque-tratamiento.

En la siguiente figura se representa una situación experimental con 4 bloques y 4 tratamientos (genotipos) en donde se observa que el ordenamiento de los tratamientos es el mismo en cada bloque (no interacción bloque-tratamiento).

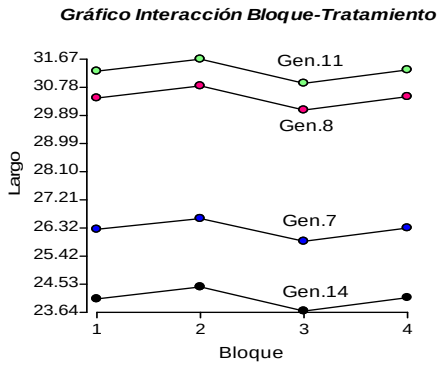


Figura 8: Gráfico para analizar el supuesto de interacción bloque-tratamiento.

Se presenta a continuación el gráfico para evaluar el supuesto de aditividad bloque-tratamiento para el modelo ajustado con los datos del archivo *Bloque*.

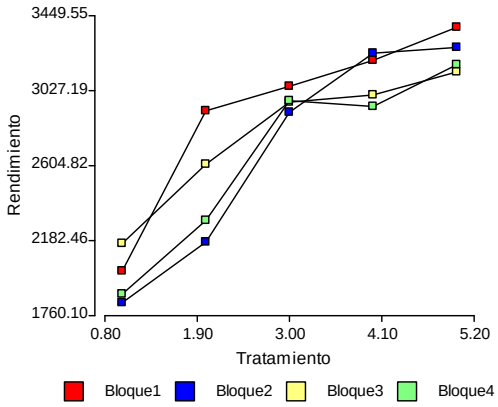


Figura 9: Gráfico para analizar el supuesto de interacción bloque-tratamiento. Archivo Bloque.

Como puede verse hay cruzamientos entre algunos perfiles, no obstante no se consideran tan graves como para dudar del cumplimiento del supuesto.

Existen pruebas formales para verificar el cumplimiento del supuesto de aditividad bloque-tratamiento (Montgomery, 1991).

Análisis de covarianza

Además de las técnicas de control de la estructura de parcelas para mejorar la precisión de las comparaciones entre tratamientos, existe otra técnica que involucra el uso de *covariables* y es denominada *Análisis de covarianza*. La covariable es una variable que se observa sobre cada una de las unidades experimentales del experimento y si se encuentra relacionada linealmente con la variable en estudio, puede ser utilizada para corregir a la variable respuesta, antes de realizar comparaciones entre tratamientos. Esta técnica es una

combinación de análisis de varianza y regresión lineal. Puede ser utilizada independientemente de la estructura de parcelas y/o tratamientos que tenga el experimento. La covariable requiere de supuestos distribucionales similares a los del análisis de regresión. Otros supuestos adicionales para el uso de esta técnica son la no interacción covariable-tratamiento y la igualdad de coeficientes de regresión entre los grupos de tratamientos. El modelo utilizado para este análisis contempla además de las estructuras de tratamientos y parcelas, la adición de una o más variables regresoras. Si el diseño es completamente aleatorizado y se tiene una covariable el modelo a ajustar es:

$$Y_{ij}=\mu+\tau_i+\beta X_{ij}+\varepsilon_{ij} \qquad \text{con } i=1,...,a;\ j=1,...,n$$

donde μ corresponde a la media general, τ_i el efecto del i-ésimo tratamiento; β es el parámetro desconocido que representa las tasa de cambio en Y frente al cambio unitario de X; X_{ij} es la variable regresora o covariable y ε_{ij} es el error aleatorio asociado con la unidad experimental. Comúnmente los términos de error se asumen normalmente distribuidos con esperanza cero y varianza común σ^2.

Si el diseño contempla alguna estructura de parcelas, estas deberán ser declaradas en el modelo como se hace habitualmente. Si se cuenta con más de una covariable, se agregarán al modelo como $\beta_1 X_{1ij}$, $\beta_2 X_{2ij}$ etc.

En la tabla de análisis de la varianza se agrega un nuevo factor (covariable) con un grado de libertad sobre el cual se pueden hacer inferencias. Las comparaciones de medias y/o contrastes a posteriori se deben realizar sobre las medias corregidas por el efecto de la covariable. Al ingresar una o más covariables en el análisis de varianza InfoStat presentará automáticamente las medias de la variable dependiente, para cada nivel de un factor de clasificación, ajustadas para los valores medios de las covariables. Para realizar análisis de los supuestos, InfoStat dará directamente los residuos y predichos ajustados por el modelo con la/s covariable/s incluidas. El residuo para un diseño a una vía de clasificación, en el caso de tener una covariable, tiene la siguiente forma:

$$e_{ij} = y_{ij} - \bar{y}_{i.} - \hat{\beta}\left(x_{ij} - \bar{x}_{i.}\right)$$

Ejemplo 21: Con el fin de comparar el incremento del diámetro a la altura del pecho (DAP) en un periodo de 5 años para tres especies de algarrobo, se realizó un estudio observacional sobre un total de 39 árboles, seleccionados al azar, de un monte en el que estaban representadas las especies nigra, flexuosa y chilensis. Una vez seleccionados los árboles a medir, se contó el número de algarrobos (sin distinción de especie) con DAP mayor a 5 cm que crecían en un radio de 15 metros (vecinos). Esta variable fue utilizada como covariable en el ANAVA para determinar diferencias en el crecimiento de las tres especies. Los datos están en el archivo Covarianza.

El archivo de datos contiene tres columnas, una identificando a la covariable ("vecinos"), otra a la especie ("Especie"), otra a la variable respuesta ("incre"). Para el análisis elegir Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA. Si en la ventana del

selector de variables de **Análisis de la varianza** se declara "Especie" como **Variable de clasificación**, "incre" como **Variable dependiente** y "vecinos" como **Covariable**, la siguiente ventana **Análisis de Varianza** señalará que las variables "Especie" y "vecinos" han sido seleccionadas como variables de clasificación y aparecerán en la subventana **Especificación de los términos del modelo**. En la solapa **Comparaciones** se eligió la prueba de LSD, en **Medias a comparar** se marcó "Especie" y se dejó el nivel de significación de 0.05 (opción por defecto). Al **Aceptar** se abrió una ventana de **Resultados** conteniendo la información que se presenta en la Tabla 36.

Tabla 36: Análisis de covarianza. Archivo Covarianza.

Análisis de la Varianza

Variable	N	R^2	R^2Aj	CV
incre	39	0.99	0.99	5.16

Cuadro de Análisis de la Varianza (SC Tipo III)

F.V.	SC	gl	CM	F	p	Coef
Modelo	1776.61	3	592.20	994.93	<0.0001	
Especie	61.70	2	30.85	51.83	<0.0001	
vecinos	1570.42	1	1570.42	2638.38	<0.0001	-4.93
Error	20.83	35	0.60			
Total	1797.45	38				

Test : LSD Fisher Alfa: 0.05 DMS: 0.63368

Error: 0.5952 gl: 35

Especie	Medias	n		
chilensis	14.06	8	A	
flexuosa	14.26	19	A	
nigra	16.84	12		B

Letras distintas indican diferencias significativas(p<=0.05)

Cuando se introducen covariables, InfoStat agrega a la tabla de ANOVA una columna con los coeficientes de regresión (Coef.) de las mismas. En este ejemplo, existe una relación lineal significativa (p<0.0001) de pendiente negativa (-4.93) entre el crecimiento y el número de vecinos. Como el número de vecinos podría ser distinto para los árboles de diferentes especies, es necesario descontar el efecto del número de vecinos sobre el crecimiento antes de realizar las comparaciones entre tratamientos (medias ajustadas por la covariable). InfoStat realiza esta operación automáticamente si se ingresa el número de vecinos como covariable en el selector de variables para este análisis.

Análisis de la varianza no paramétrico

Prueba de Kruskal-Wallis

Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA NO PARAMÉTRICO ⇒ KRUSKAL-WALLIS. Permite realizar un análisis de varianza no paramétrico a una vía de clasificación. El ANAVA propuesto por Kruskal y Wallis (1952) permite comparar las esperanzas de 2 o más distribuciones sin necesidad de realizar el supuesto de que los términos de error se distribuyen normalmente.

La hipótesis nula establece que $\mu_1=\mu_2=,...,=\mu_a$, donde μ_i representa la esperanza del i-ésimo tratamiento, con *i*=1, 2,....,*a*. Esta prueba se aplica cuando se tienen muestras independientes de cada población, con observaciones de naturaleza continua y las varianzas poblacionales son iguales.

El estadístico de la prueba (*H*) se basa en la suma de los rangos asignados a las observaciones dentro de cada tratamiento. Su distribución exacta es obtenida a partir de la consideración de todas las configuraciones posibles de los rangos de *N* observaciones en *a* grupos de n_i observaciones cada uno.

InfoStat utiliza la distribución exacta del estadístico para casos donde las configuraciones totales de los rangos no es superior a 100000. El número de configuraciones posibles de los rangos crece rápidamente con el incremento del número de tratamientos y/o el número de repeticiones por tratamiento. Para tres tratamientos con tres repeticiones cada uno, el número de configuraciones es 1680 pero si cada tratamiento tiene 5 repeticiones el número de configuraciones ya es mayor a 100000. Para situaciones experimentales donde el número de configuraciones es mayor a 100000 InfoStat obtiene los valores *p* de la prueba a través de la aproximación de la distribución del estadístico a la distribución Chi cuadrado, con *a*-1 grados de libertad.

Para realizar el ANAVA no paramétrico, en la ventana **Prueba de Kruskal-Wallis**, indicar en **Variables** la/s variable/s que actuarán como dependiente/s y, como **Criterios de clasificación,** los factores a considerar, es decir, la o las columnas del archivo que definen los grupos para los cuales se comparan las medias. El usuario puede requerir medias, medianas, desvío estándar (DE), rangos promedio y tamaño muestral (*N*) para cada tratamiento. Además se puede requerir el estadístico de la prueba de Kruskal Wallis (*H*), el valor *p* asociado (*p*) y el factor de corrección que InfoStat utiliza para actualizar el estadístico en casos de empates (*C*).

InfoStat permite solicitar **Comparaciones de a pares** entre las medias de los rangos de tratamientos y/o **Contrastes** entre medias de los rangos de tratamientos. El procedimiento usado para juzgar la significancia de las comparaciones múltiples y los contrastes postulados es el descripto en Conover (1999). Si bien las comparaciones entre tratamientos se realizan a través de las diferencias entre las medias de los rangos, InfoStat permite también visualizar las diferencias entre tratamientos a nivel de medias y medianas de los valores originales de las variables.

Prueba de Friedman

Menú ⇒ ESTADÍSTICAS ⇒ ANÁLISIS DE LA VARIANZA NO PARAMÉTRICO ⇒ FRIEDMAN, permite realizar un análisis de varianza no paramétrico a dos vías de clasificación. El ANAVA propuesto por Friedman (1937, 1940) permite comparar las esperanzas de 2 o más distribuciones cuando el diseño de la experiencia ha sido en bloques completos aleatorizados, sin necesidad de verificar el cumplimiento del supuesto de normalidad.

Esta prueba requiere que las observaciones sean independientes y que las varianzas poblacionales sean homogéneas. La hipótesis a probar es: H_0: $\mu_1=\mu_2=,...,=\mu_a$

donde μ_i representa la esperanza del i-ésimo tratamiento; con *i*=1, 2,....,*a.*

Ejemplo 22: Supóngase que en 12 hogares se prueba la durabilidad de 4 tipos de bombillas eléctricas. Cada jefe de hogar reportará la duración en horas de cada tipo de bombilla. Los datos se encuentra en el archivo Friedman. Las hipótesis que se someten a prueba son H_0: No hay diferencias de durabilidad entre los cuatro tipos versus H_1: Al menos un tipo tiene diferente durabilidad.

El diseño de esta experiencia se podría asimilar a un diseño en bloques completos aleatorizados. Las observaciones de cada jefe de hogar podrían considerarse más correlacionadas que las observaciones provenientes de distintos jefes, por provenir de la pruebas realizadas en el mismo hogar. Cada hogar puede ser considerado como un *bloque*. La estrategia de bloqueo de la observaciones se realiza a los fines de distinguir la variación debida a los hogares (los cuales probablemente realicen un uso diferente de las bombillas), de la variación aleatoria o error experimental. Cada hogar evalúa los 4 tipos o *tratamientos*. Para realizar la prueba de Friedman se aplica la transformación *rango* a la variable en estudio dentro de cada hogar o bloque.

Observación: para poder realizar esta prueba InfoStat requiere de un archivo con *a* columnas, una para cada tratamiento y cada "caso" corresponde a un bloque.

En la ventana **Prueba de Friedman** se debe indicar cuales son las columnas del archivo que representan los tratamientos en **Variables que definen tratamientos**, en el ejemplo son: Bombilla 1, Bombilla 2, Bombilla 3 y Bombilla 4. En la siguiente ventana se pueden requerir comparaciones múltiples "a posteriori" con el nivel de significación deseado. Los procedimientos de comparación son basados en las medias de los rangos por tratamiento y en la varianza de los rangos según se describe en Conover (1999).

Tabla 37: Prueba de Friedman.

Prueba de Friedman

Bombilla 1	Bombilla 2	Bombilla 3	Bombilla 4	T²	p
3.21	1.96	2.04	2.79	3.28	0.0331

Minima diferencia significativa entre suma de rangos (11.498)

Tratamiento	Suma(Ranks)	Media(Ranks)	n			
Bombilla 2	23.50	1.96	12	A		
Bombilla 3	24.50	2.04	12	A	B	
Bombilla 4	33.50	2.79	12	A	B	C
Bombilla 1	38.50	3.21	12			C

Letras distintas indican diferencias significativas(p<= 0.050)

Análisis de regresión lineal

Menú ⇒ ESTADÍSTICAS ⇒ REGRESIÓN LINEAL, permite estudiar la relación funcional entre una variable respuesta Y (variable dependiente) y una o más variables regresoras X (variables independientes o predictoras). El primer caso se conoce como Regresión Lineal Simple y el segundo como Regresión Lineal Múltiple (Draper y Smith, 1998).

Mediante la regresión se estudia cómo los cambios en la/s variable/s predictora/s afectan a la variable respuesta, mediante el ajuste de un modelo para la relación funcional entre ambas. Genéricamente, la relación entre las variables se modela de la forma $\mathbf{Y}=\mathbf{X}\beta+\boldsymbol{\varepsilon}$, donde **Y** es el vector de observaciones, **X** es la matriz que contiene a las variables regresoras, **ß** es un vector de parámetros que serán estimados a partir de los datos y $\boldsymbol{\varepsilon}$ es el vector de términos de error aleatorios.

InfoStat usa el método de mínimos cuadrados para obtener estimaciones de los coeficientes de la ecuación que explica la relación entre las variables. A partir de estos coeficientes se construye la ecuación de predicción que permite conocer el valor predicho de Y para cualquier valor de la/s variable/s regresora/s dentro del dominio de los valores experimentados. InfoStat también realiza el análisis de regresión por mínimos cuadrados ponderados para considerar situaciones de heterogeneidad de varianzas de los términos de error.

A través del Análisis de la Varianza se puede conocer cuánto de la variación de los datos es explicada por la regresión y cuánto debe considerarse como no explicada o residual. Si la variación explicada es sustancialmente mayor que la variación no explicada, el modelo propuesto será bueno para fines predictivos. Una medida de la capacidad predictiva del modelo es el coeficiente de determinación R^2 que relaciona la variación explicada por el modelo con la variación total.

Cuando los datos contienen más de una observación para al menos un valor de la variable independiente, se puede obtener una medida del error puro (no ocasionado por la mala especificación del modelo) y se puede probar la falta de ajuste del modelo (falta de ajuste).

Para identificar un buen modelo y para controlar el cumplimiento de los supuestos del análisis, InfoStat provee diferentes medidas consideradas como criterios de diagnóstico.

Modelo

La ecuación del modelo de regresión lineal múltiple es:

$$Y_i=\beta_0 + \beta_1 x_{1i} +\beta_2 x_{2i} + \ldots..+ \beta_k x_{ki} + \varepsilon_i$$

donde,

Y_i= i-ésima observación de la variable dependiente Y

$x_{1i}, x_{2i},...,x_{ki}$ = i-ésimo valor de las variables regresoras $X_1, X_2,...,X_k$ o independientes

β_0 = parámetro desconocido que representa la ordenada al origen de la recta (indica el valor esperado de Y cuando $x_1 = 0$, $x_2 = 0,...,x_k = 0$)

β_1, ...,β_k = parámetros desconocidos que representan las tasas de cambio en Y frente al cambio unitario de $X_1,X_2...,X_k$, respectivamente

ε_i = término de error aleatorio

En general, se supone que dentro del dominio estudiado, la relación entre la respuesta y las predictoras es bien aproximada por el modelo de regresión lineal propuesto.

Las variables predictoras son variables seleccionadas por el investigador que representan a las poblaciones a partir de las cuales fueron obtenidas las respuestas.

Los errores aleatorios son independientes y normalmente distribuidos con media cero y varianza constante.

Menú ⇒ ESTADÍSTICAS ⇒ REGRESIÓN LINEAL, permite realizar este análisis donde las variables involucradas son declaradas en la ventana **Análisis de regresión lineal**. La variable Y se debe colocar como **Variable dependiente** y las variables X se asignan como **Variables regresoras**. La subventana **Pesos** deberá ser utilizada cuando se desean ponderar las observaciones de la variable respuesta en forma diferencial. Si no se indica ninguna columna del archivo como aquella que contiene los pesos para realizar dicha ponderación, las estimaciones reportadas para los parámetros del modelo son obtenidas por el método de estimación por mínimos cuadrados ordinarios. Si se indican pesos las estimaciones reportadas son obtenidas por el método de mínimos cuadrados ponderados.

Para realizar análisis por grupos, la variable que separa los datos en grupos se declara en la solapa **Particiones** opción **Seleccionar por**. En esta solapa también se encuentra la opción **Peso**, para indicar la variable que representa a los pesos en un análisis por mínimos cuadrados ponderados.

La ventana de la prueba muestra las solapas **General, Diagnóstico, Polinomios, Hipótesis** y **Selección de modelo**.

Solapa **General**: permite seleccionar la información que se mostrará en los resultados. Por defecto en los resultados se muestran la matriz de coeficientes de regresión y la tabla del análisis de la varianza, pero se pueden agregar criterios de diagnóstico y la matriz de covarianza de los coeficientes de regresión. InfoStat permite seleccionar las siguientes opciones:

Coeficientes de regresión y estadísticos asociados: reporta para cada parámetro incluido en el modelo (Coef.) el valor estimado (Est), el error estándar de la estimación (E.E.), los límites del intervalo de confianza al 95% (LI y LS), el valor del estadístico *T* para probar la hipótesis que el parámetro vale cero, el valor de significación *p* para la prueba de hipótesis basada en *T* y el índice *Cp de Mallows.*

El **Cp de Mallows** : Para cada término del modelo InfoStat calcula el índice *Cp* de la siguiente manera:

$$Cp = \left({SCError_p} \Big/ {CMError} \right) - \left(n - 2p \right)$$

donde $SCError_p$ es la suma de cuadrados del error de un modelo reducido (con *p* parámetros incluida la constante) respecto al modelo completo especificado por el usuario. El modelo reducido contiene todos los términos del modelo completo menos el término de la línea en la que se reporta el *Cp*. El valor CMError es el cuadrado medio de error para el modelo completo especificado por el usuario y *n* el número total de observaciones. Luego, para cada regresora se tiene un indicador de su contribución en el ajuste del modelo propuesto por el usuario, ya que valores de *Cp* cercanos a *p* corresponden a modelos con pequeño sesgo en la predicción. Si al sacar una regresora el valor *Cp* se incrementa mucho se puede pensar que esa regresora es importante para el ajuste del modelo.

En los casos de regresión múltiple la eliminación de una o más variables regresoras puede incrementar el valor predictivo del modelo aún cuando el R^2 disminuye. El *Cp* de Mallows para el término constante no es reportado por InfoStat.

Tabla del análisis de la varianza: muestra el coeficiente R^2, el R^2 ajustado, el error cuadrático medio de predicción y el análisis de la varianza para el modelo especificado. Esta tabla incluirá la prueba para el ajuste del modelo (reportada como *Lack of fit*) cuando en la subventana **Opciones,** se elija la opción **Error puro**.

El coeficiente R^2 mide la proporción de la variación en Y que es explicada por la relación con X. El R^2 se calcula como el cociente entre la suma de cuadrados del modelo y la suma de cuadrados total. El R^2 ajustado se obtiene a partir de la expresión:

$$R^2_{Aj} = 1 - (1 - R^2)\left[(n-1)/(n-p)\right]$$

donde *n* es el total de observaciones y *p* el número de parámetros del modelo ajustado.

En el análisis de la varianza, por defecto las sumas de cuadrados son de tipo III, pero se puede indicar el uso de la suma de cuadrados tipo I. Las sumas de cuadrados tipo I se llaman sumas de cuadrados secuenciales porque particionan la suma de cuadrados del modelo según la secuencia de incorporación de términos al modelo. Es decir, suponiendo que el modelo especificado es $Y= \beta_0+\beta_1X_1+\beta_2X_2$, la suma de cuadrados para X_1 es la suma de cuadrados para el modelo que contiene la constante y X_1, la suma de cuadrados para X_2 es la reducción de la suma de cuadrados del error debida a que se incorpora al modelo X_2. Como las sumas de cuadrados tipo I son dependientes del orden en que los términos se adicionan al modelo, bajo ordenamientos diferentes se tendrán diferentes sumas de cuadrados. InfoStat calculará las sumas de cuadrados tipo I incorporando los términos al modelo según el orden asignado a las variables en el selector de variables. Las sumas de cuadrados tipo I son especialmente recomendadas para modelos de regresión polinómicos.

Tabla resumen de criterios de diagnóstico: muestra los valores extremos, máximo (max) y mínimo (min), de los residuos estudentizados (rs), los residuos externamente

estudentizados (res), el leverage (Lev) y la distancia de Cook (Cook), identificando el caso al que están asociados cada uno de estos valores extremos.

Matriz de covarianza: muestra la matriz de covarianza de los estimadores de los coeficientes de regresión.

Subventana **Opciones:** permite adicionar al modelo la **ordenada al origen**, especificar que se trabaje con **regresoras centradas** (previo al análisis de regresión, las regresoras son centradas por su media. El término constante corresponde a la respuesta media para las condiciones medias de las regresoras), solicitar la prueba de **Atkinson**, obtener el cálculo del **error puro** para probar la falta de ajuste del modelo propuesto y obtener todas las **regresiones simples**.

La prueba de Atkinson permite establecer si es necesario usar la transformación de potencia:

$$Y^{*} = \left(Y^{\lambda} - 1 \right) \Big/ \lambda$$

InfoStat estima un parámetro γ que se relaciona con la transformación de potencia como: $\lambda=1-\gamma$. La prueba de Atkinson permite contrastar la hipótesis $\gamma=0$. Si la hipótesis no se rechaza implica que $\lambda=1$ y por lo tanto no es necesario aplicar la transformación de potencia a los datos. Si la prueba resulta significativa la transformación de potencia es aconsejable y el exponente de la transformación está dado por $1-\gamma$. La estimación de γ se reporta en la Tabla de coeficientes de regresión como el valor estimado (Est) para el coeficiente de Atkinson. Esta prueba no se puede realizar si la variable independiente tiene valores cero. En este caso se puede transformar la variable sumando una constante para obtener los resultados de la prueba.

La prueba de bondad de ajuste (falta de ajuste) permite corroborar si el modelo utilizado ajusta bien. Requiere una estimación de σ^2 independiente del modelo, que se llama "error puro". Para estimarlo hace falta tener más de una observación para al menos un punto del dominio de la variable regresora. Los resultados se leen en la tabla de análisis de la varianza.

Solapa **Diagnóstico**: permite seleccionar elementos de diagnóstico, indicar el cálculo de los valores predichos y solicitar los intervalos de confianza y de predicción, eligiendo un nivel de confianza. La información solicitada se guardará como nuevas variables en el archivo original. El cálculo de los valores predichos, y de los intervalos de confianza y de predicción también se realizará para aquellos valores de X, en la tabla de datos, que no tuviesen el correspondiente valor de Y. De la misma forma, si se deseara conocer el valor predicho usando un valor *x* que no esté en el conjunto de datos se deberá ingresar dicho valor en la tabla y correr nuevamente el análisis solicitando los cálculos de interés.

Los valores **predichos** son los valores de la variable dependiente que se obtienen usando el modelo ajustado. El modelo ajustado se construye con las estimaciones de los parámetros.

Los **intervalos de confianza** son los intervalos para la esperanza de Y dado X=x_0, que en el caso de la regresión simple están dados por la siguiente expresión:

$$\hat{y}_0 \pm T_{1-\alpha/2}\sqrt{S^2\left[\frac{1}{n}+\frac{\left(x_0-\overline{x}\right)^2}{\sum x_i^2-\left(\sum x_i\right)^2/n}\right]}$$

donde $T_{1-\alpha/2}$ es el cuantil correspondiente en una distribución *T* de Student con *n*-2 grados de libertad y S^2 es la estimación de σ^2. Cuando los intervalos de confianza se obtienen para todos los valores de X en un recorrido dado y se unen, se obtienen las **bandas de confianza**.

Los **intervalos de predicción** son intervalos para los valores de Y dado **x**=$\mathbf{x}_0$, que en el caso de la regresión simple tienen la siguiente expresión:

$$\hat{y}_0 \pm T_{1-\alpha/2}\sqrt{S^2\left[1+\frac{1}{n}+\frac{\left(x_0-\overline{x}\right)^2}{\sum x_i^2-\left(\sum x_i\right)^2/n}\right]}$$

En el caso de regresión múltiple la expresión para este intervalo será:

$$\hat{y}_0 \pm T_{1-\alpha/2}\sqrt{S^2\left[1+\frac{1}{n}+\left(\mathbf{x}_0-\overline{\mathbf{x}}\right)^T\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}\left(\mathbf{x}_0-\overline{\mathbf{x}}\right)\right]}$$

donde $\hat{y}_0$ es el valor predicho, $\mathbf{x}_0$ es el vector de valores dados para la predicción deseada, $\overline{\mathbf{x}}$ es el vector de medias de las variables regresoras y **X** es la matriz de diseño. El cuantil $T_{1-\alpha/2}$ corresponde a una distribución *T* de Student con *n-p* grados de libertad, donde *p* es el número de parámetros (coeficientes del modelo de regresión).

Cuando se obtienen los intervalos de predicción de Y para los valores observados en la muestra y se unen entre sí, los límites superiores y entre sí los inferiores, se obtienen las **bandas de predicción**.

La diferencia entre intervalo de confianza y de predicción radica en que el primero delimita una región, que con probabilidad 1-α contiene a la esperanza de Y dadas las regresoras, mientras que los límites de un intervalo de predicción son aquellos dentro de los cuales se espera, con probabilidad 1-α, observar una realización futura de Y dadas las regresoras.

En la solapa **Diagnóstico** se pueden activar los campos **Graficar ajuste, bandas de confianza y predicción** para obtener en forma automática un gráfico con estos componentes.

Como medidas de diagnóstico, InfoStat permite obtener los *residuos* (RDUO_"nombre de la variable"), *residuos estudentizados* (RE_"nombre de la variable"), *residuos externamente estudentizados* (REE_"nombre de la variable"), *distancia de Cook* (COOK_"nombre de la variable"), *leverage* (LEVE_"nombre de la variable") y *residuos parciales* (RPAR_"nombre

de la variable"), que son adicionados a la tabla de datos (bajo el nombre que se muestra entre paréntesis) y se calculan como:

Residuo estudentizado: está dado por el cociente entre el residuo asociado a la observación *i* y la raíz cuadrada del producto entre el cuadrado medio del error (S^2) y el término ($1-h_{ii}$), donde h_{ii} es el leverage.

$$RE = e_i \Big/ \sqrt{S^2\left(1-h_{ii}\right)}$$

Residuo externamente estudentizado: es el cociente entre el residuo asociado a la observación *i* y la raíz cuadrada del producto entre el cuadrado medio del error (S_i^2), calculado después de la eliminación del residuo *i*, y el término ($1-h_{ii}$), donde h_{ii} es el *leverage*.

$$REE = e_i \Big/ \sqrt{S_i^2\left(1-h_{ii}\right)}$$

Leverage: es una medida de la contribución de la observación i-ésima al i-ésimo valor ajustado. Los *leverage* son los elementos diagonales (h_{ii}) de la matriz **H**, con $\mathbf{H=X(X'X)^{-1}X'}$ donde **X** es la matriz de diseño conformada por las variables regresoras. El producto **HY** origina los valores ajustados siendo **Y** el vector de observaciones.

Distancia de Cook: permite medir la influencia de la i-ésima observación. Grandes valores de esta medida indican observaciones cuya eliminación tiene gran influencia sobre los valores predichos. La expresión para su cálculo es:

$$Cook = \left(e_i \Big/ \sqrt{S^2\left(1-h_{ii}\right)} \right)^2 \left(h_{ii} \Big/ 1-h_{ii} \right)\left(1 \Big/ p \right)$$

donde *p* es el número de parámetros del modelo.

Residuos parciales: son obtenidos, en el marco de una regresión múltiple, a partir de los residuos asociados al modelo completo más el producto de una regresora particular y su coeficiente de regresión. Por ejemplo, en una regresión múltiple que involucra dos regresoras X_1 y X_2, el residuo parcial asociado a X_2 para la observación i-ésima se obtiene de la siguiente forma: 1) ajustando el modelo $Y_i=\beta_0 +\beta_1 X_{1i} +\beta_2 X_{2i} + \varepsilon_i$ y obteniendo el residuo e_i 2) calculando $RP_i= e_i+\beta_2 X_{2i} + \varepsilon_i$

El botón **Típico** activa como elementos de diagnóstico a aquellos estadísticos típicamente seleccionados en la etapa de diagnóstico y evaluación del modelo ajustado. Estos son: Residuos estudentizados, predichos y leverage.

Si se guardan elementos de diagnóstico, estos serán adicionados como nuevas columnas a la tabla activa, de modo que si se realiza el análisis varias veces, se generarán tantas nuevas

columnas como elementos se pidan (y serán numerados consecutivamente). Para evitar esta proliferación de columnas se puede activar el campo **Sobreescribir**.

El análisis de los estadísticos de diagnóstico se puede completar mediante gráficos. Un gráfico comúnmente usado en regresión lineal es el *diagrama de dispersión.* Usualmente se realiza más de un diagrama: dispersión de Y versus X (o cada X en el caso de regresión múltiple), dispersión de residuos versus predichos, residuos versus X, leverage y/o residuos parciales versus cada una de las regresoras. Mediante estos gráficos se pueden detectar valores extremos, puntos influyentes, violaciones de supuestos y el tipo de relación que hay entre las variables consideradas para mejorar el modelo propuesto. Para obtener estos gráficos en la solapa **Diagnóstico** se deben activar los campos correspondientes en **Guardar.** Para construir otros gráficos ver el Capítulo Gráficos.

Solapa **Polinomios**: permite hacer un ajuste polinómico de grado *n,* incluyendo en el modelo términos cuadráticos, cúbicos y de orden *n*, indicado por el usuario para las variables predictoras que se seleccionen.

Solapa **Hipótesis:** permite especificar hipótesis sobre uno o más coeficientes del modelo de regresión bajo la forma general **Hb=h**. El **vector b** es el vector de coeficientes del modelo de regresión y **H** es una matriz de contrastes. Las filas de **H** contendrán los coeficientes (especificados por el usuario) de una combinación lineal de los elementos de **b** y **h** es el vector que contendrá los valores hipotéticos de **Hb**, especificados por el usuario. Si no se ingresan valores para **h,** el sistema asumirá que los valores son ceros.

Solapa **Selección de modelo:** permite implementar la **eliminación *backward*** indicando el máximo valor-p para retener la componente en el modelo.

Ejemplo 23: Para estudiar la relación entre la biomasa y el pH en un medio de cultivo, se midió la biomasa (gr) para valores de pH entre 3 y 7 registrándose 45 mediciones. Los datos se encuentran en el archivo RegLin.

Se tomó como **Variable dependiente** a la biomasa y como **Variable regresora** al pH.El siguiente gráfico, obtenido por defecto, muestra el comportamiento de las variables.

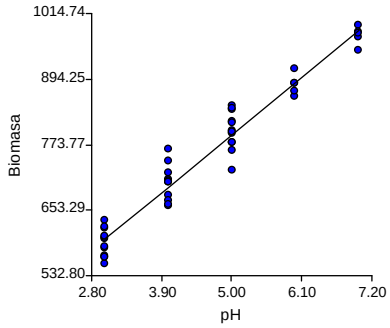


Figura 10: Diagrama de dispersión. Archivo RegLin.

El diagrama indicaría que hay una relación positiva entre la biomasa y el pH. El análisis de regresión reportó los siguientes resultados:

Tabla 38: Análisis de regresión lineal. Archivo Reglin.

Análisis de regresión lineal

Variable	N	R^2	R^2Aj
Biomasa	45	0.95	0.95

Coeficientes de regresión y estadísticos asociados

Coef.	Est.	E.E.	LI(95%)	LS(95%)	T	Valor p	CpMallows
const	313.95	15.87	281.94	345.95	19.78	<0.0001	
pH	95.56	3.35	88.80	102.32	28.51	<0.0001	795.40

Tabla de análisis de la varianza SC Tipo III

FV	SC	gl	CM	F	Valor p
Modelo	685884.36	1	685884.36	812.85	<0.0001
pH	685884.36	1	685884.36	812.85	<0.0001
Error	36283.65	43	843.81		
Lack of fit	2367.38	3	789.13	0.93	0.4348
Error puro	33916.27	40	847.91		
Total	722168.01	44			

Como puede verse, en la tabla del análisis de la varianza, hay relación lineal entre la biomasa y el pH (p<0.0001). También se observa que el modelo propuesto no presenta falta de ajuste (p=0.4348). Tomando la información sobre los coeficientes de regresión se puede escribir la ecuación del modelo ajustado:

$$\hat{y} = a+bx = 313.95 + 95.56x$$

Esta recta permite estimar el valor de y (valor predicho) para un valor de x. El modelo ajustado puede ser usado con fines predictivos; por ejemplo, para un pH de 3.5 la biomasa esperada es $\hat{y} = 313.95 + 95.56(3.5) = 648.41 gr$. Este resultado, al igual que cualquier otra predicción deseada usando valores de X dentro o fuera del rango estudiado, puede obtenerse automáticamente al ingresar como dato en la columna *pH* el valor 3.5 y correr de nuevo el análisis pidiendo el cálculo de predichos. El siguiente gráfico fue obtenido al solicitar en la solapa **Dignóstico** la obtención de las **bandas de confianza** y **de predicción**:

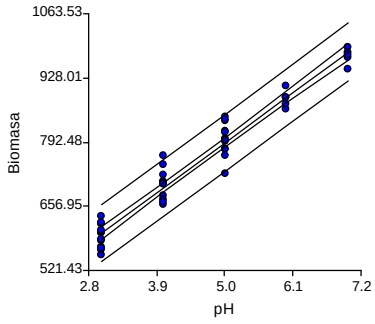


Figura 11: Diagrama de dispersión con bandas de confianza y de predicción. Archivo RegLin.

En la figura anterior la línea central corresponde al modelo ajustado, las líneas siguientes corresponden a las bandas de confianza y las líneas externas a las bandas de predicción.

Para visualizar algunos elementos de diagnóstico se solicitaron los gráficos de residuos estudentizados vs. predichos, leverage y distancia de Cook. InfoStat generó los siguientes:

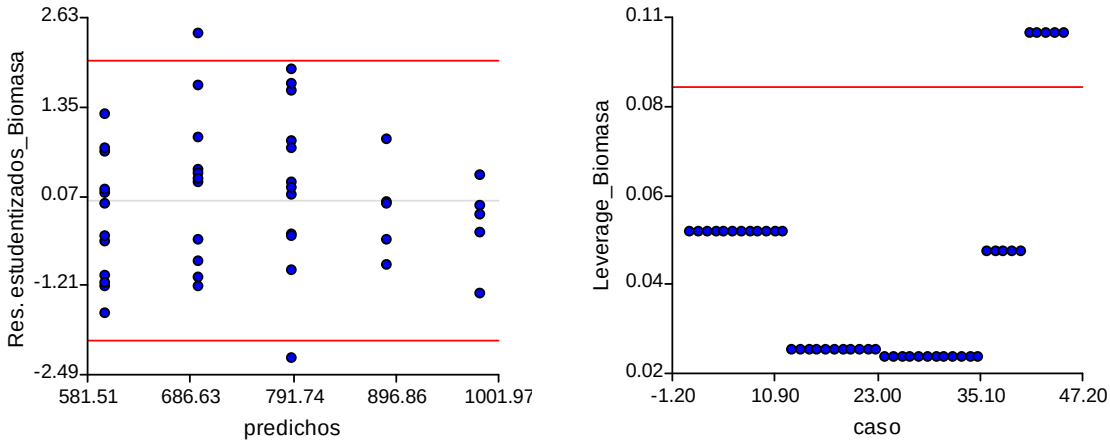


Figura 12: Gráfico de los residuos estudentizados y valores predichos y Gráfico de los leverages. Archivo RegLin.

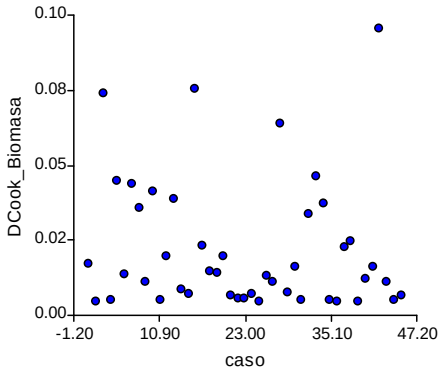


Figura 13: Gráfico de las distancias de Cook. Archivo RegLin.

Validación de los supuestos

Normalidad

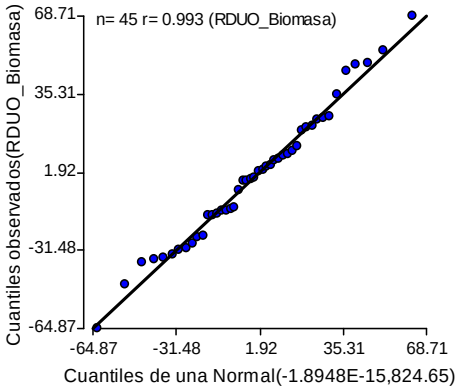


Figura 14: Gráfico Q-Q plot. Archivo RegLin.

Obsérvese que en el Q-Q plot fue realizado con los residuos del modelo de regresión y usando como distribución teórica la Normal (ver QQ-plot). Los puntos se disponen en una recta a 45º indicando que el supuesto distribucional para los residuos se cumple. Realizando la prueba se Shapiro-Wilks (modificada) en el menú INFERENCIA BASADA EN UNA MUESTRA se concluye que los datos siguen una distribución normal (p=0.8295). Los resultados de esta prueba se presentan en la siguiente tabla.

Tabla 39: Prueba de Shapiro-Wilks para los residuos del análisis de regresión lineal. Archivo Reglin.

Shapiro-Wilks (modificado)

Variable	n	Media	D.E.	W*	p (una cola)
RDUO_Biomasa	45	0.00	28.72	0.98	0.8292

Homocedasticidad

En Figura 15, puede verse que los puntos para los valores de pH más altos presentan menor dispersión que el resto, razón por la cual una prueba formal de homogeneidad de varianzas sería recomendable.

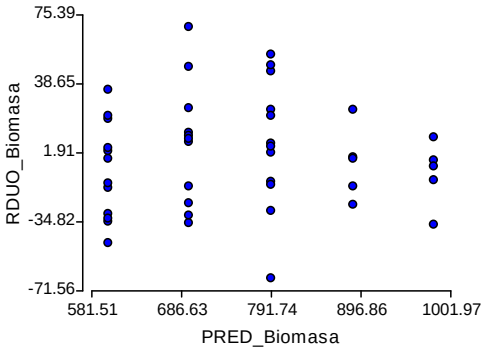
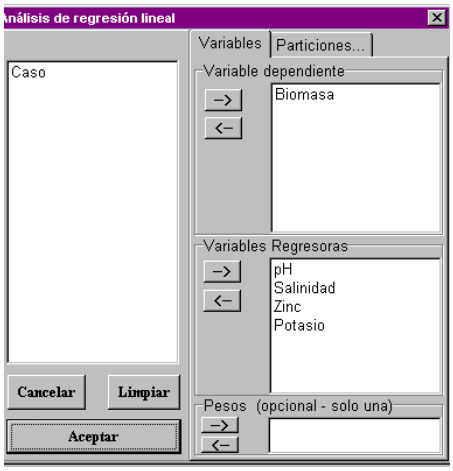


Figura 15: Gráfico de residuos versus predichos. Archivo RegLin.

A continuación se presenta un ejemplo de aplicación de la técnica de regresión lineal múltiple.

Ejemplo 24: Para estudiar la relación entre el pH (pH), la salinidad (Salinidad), el contenido de Zn (Zinc) y el contenido de K (Potasio) presentes en el suelo con la producción de biomasa de una forrajera, se registraron 45 mediciones de la biomasa (gr) y de dichos valores característicos del suelo donde crecieron las plantas. Los datos se encuentran en el archivo Salinidad.



En el selector de variables, se indicó como **Variable dependiente** a la "Biomasa" y como **Variables regresoras** a "pH", "Salinidad", "Zinc" y "Potasio".

Posteriormente, en la ventana **Análisis de regresión lineal**, se indicó qué información se deseaba visualizar en la ventana de resultados.

En este ejemplo se seleccionó la tabla de análisis de varianza asociada al modelo de regresión lineal y la tabla conteniendo las estimaciones, errores estándares y otros estadísticos asociados a cada coeficiente del modelo de regresión propuesto (definido a partir de las variables regresoras seleccionadas). También se podría pedir una tabla de resumen de criterios de diagnóstico para la evaluación del modelo ajustado. En la subventana **Opciones** es posible indicar si el modelo debe incluir o no un término correspondiente a la **Ordenada al origen**, si se desea trabajar con las **Regresoras centradas** por su media, si la tabla de análisis de varianza debe contener la prueba propuesta por **Atkinson**, si se desea una estimación del **Error puro** y si la salida debe contener **Todas las regresiones simples** que se pueden ajustar a partir del conjunto de regresoras seleccionadas.

En la solapa **Diagnóstico**, se pueden indicar las medidas de diagnóstico que se desean visualizar. Si se presiona el botón Típico se guardarán en la tabla de datos activa los **Residuos estudentizados**, los valores **Predichos** por el modelo ajustado y los **Leverage** de cada observación. El campo **Graficar residuos parciales** permite obtener automáticamente gráficas de residuos parciales para cada regresora del modelo propuesto. Este campo debiera activarse en un primer paso en un análisis de regresión lineal múltiple ya que permite tener una idea preliminar de la adecuación del modelo. En la Figura 16 se presentan estos gráficos, en los que se puede observar que: 1) Existe una relación lineal positiva entre la biomasa y el pH del suelo; 2) Existe una relación lineal negativa entre biomasa y el

contenido de sal del suelo, pero el gráfico sugiere además la presencia de una componente cuadrática en esta relación; 3) Existe una relación lineal negativa entre biomasa y el contenido de zinc del suelo; 4) Aparentemente no existe relación lineal entre biomasa y el contenido de potasio del suelo.

En la Tabla 40**,** se presenta el resultado que se obtuvo al ajustar el modelo de regresión con todas las variables regresoras.

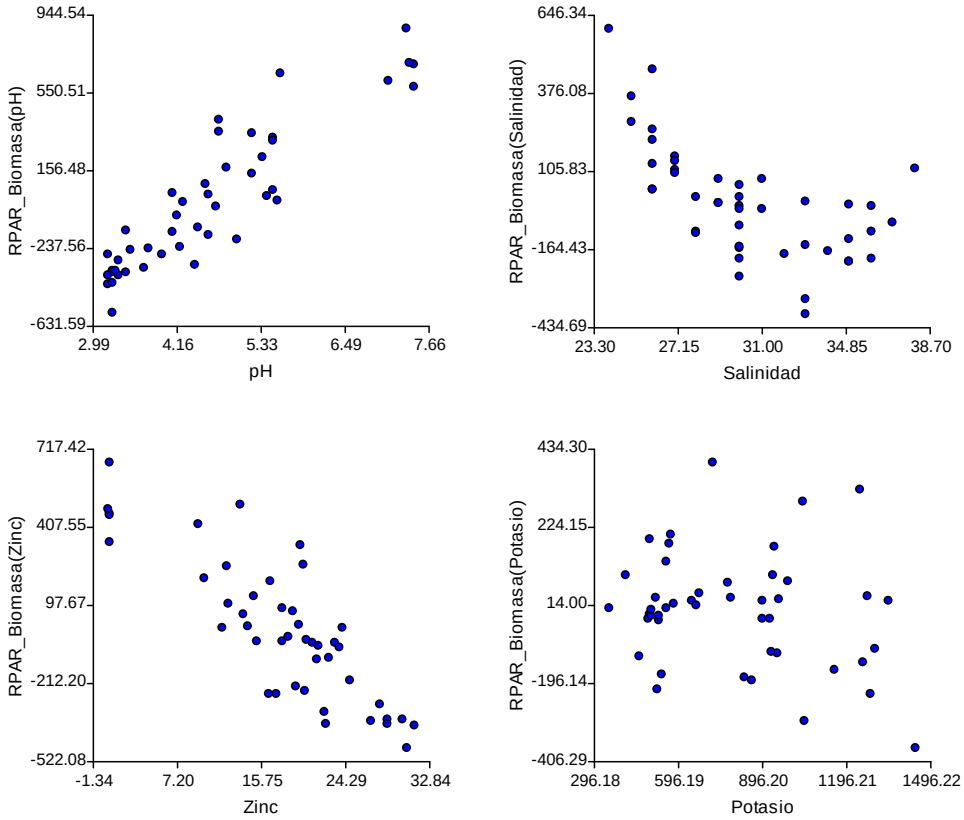


Figura 16: Gráficos de residuos parciales de biomasa para cada una de las variables regresoras. Archivo Salinidad.

Tabla 40: Análisis de regresión lineal para el modelo de regresión lineal múltiple. Archivo Salinidad.

Análisis de regresión lineal

Variable	N	R^2	R^2Aj
Biomasa	45	0.92	0.92

Coeficientes de regresión y estadísticos asociados

Coef.	Est.	E.E.	LI(95%)	LS(95%)	T	Valor p	CpMallows
const	1492.81	453.60	576.05	2409.57	3.29	0.0021	
pH	262.88	33.73	194.71	331.05	7.79	<0.0001	63.28
Salinidad	-33.50	8.65	-50.99	-16.01	-3.87	0.0004	18.65
Zinc	-28.97	5.66	-40.42	-17.52	-5.11	<0.0001	29.55
Potasio	-0.12	0.08	-0.28	0.05	-1.40	0.1680	5.95

Tabla de análisis de la varianza SC Tipo III

FV	SC	gl	CM	F	Valor p
Modelo	12120944.19	4	3030236.05	120.01	<0.0001
pH	1533665.03	1	1533665.03	60.74	<0.0001
Salinidad	378485.90	1	378485.90	14.99	0.0004
Zinc	660588.37	1	660588.37	26.16	<0.0001
Potasio	49785.48	1	49785.48	1.97	0.1680
Error	1009974.02	40	25249.35		
Total	13130918.21	44			

Se puede ver que pH, Salinidad y Zinc presentaron un valor $p<0.05$, es decir presentan relación lineal significativa. Potasio presentó un valor $p=0.1680$, es decir no fue significativa su relación lineal con biomasa.

En el análisis de regresión múltiple no se deberían eliminar variables regresoras del modelo sin antes asegurar la adecuación de éste. Por este motivo, antes de eliminar Potasio como regresora, se ajustó un nuevo modelo incorporando la componente cuadrática para la regresora Salinidad, siguiendo los mismos pasos usados anteriormente para ajustar el modelo, y agregando en la solapa **Polinomios** un polinomio de grado 2 para Salinidad como sugería el grafico de residuos parciales de biomasa versus Salinidad. Los resultados obtenidos con este nuevo modelo se presentan en la Tabla 41. Cuando a una o más variables regresoras se le incorpora uno o más términos polinómicos, InfoStat presentará automáticamente en la ventana de resultados la tabla de análisis de varianza para regresión con las sumas de cuadrados de tipo I (secuenciales) y la tabla con las sumas de cuadrado de tipo III.

Tabla 41: Análisis de regresión lineal para el modelo de regresión lineal múltiple, con la adición de un término cuadrático para la regresora salinidad. Archivo Salinidad.

Análisis de regresión lineal

Variable	N	R^2	R^2Aj
Biomasa	45	0.97	0.96

Coeficientes de regresión y estadísticos asociados

Coef.	Est.	E.E.	LI(95%)	LS(95%)	T	Valor p	CpMallows
const	10430.36	1327.07	7746.11	13114.61	7.86	<0.0001	
pH	224.02	23.56	176.37	271.68	9.51	<0.0001	93.19
Zinc	-36.39	3.99	-44.46	-28.32	-9.12	<0.0001	86.17
Potasio	-0.17	0.06	-0.28	-0.06	-3.02	0.0044	13.94
Salinidad	-590.47	80.66	-753.62	-427.32	-7.32	<0.0001	57.27
Salinidad^2	8.90	1.29	6.30	11.50	6.92	<0.0001	51.76

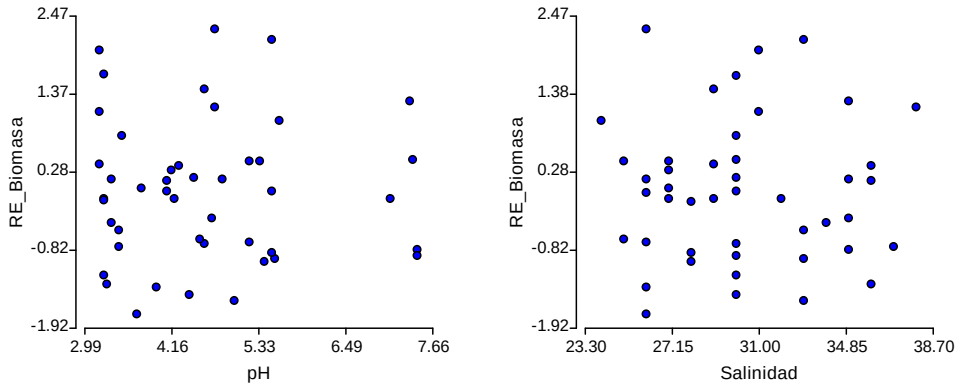
Tabla de análisis de la varianza SC Tipo I

FV	SC	gl	CM	F	Valor p
Modelo	12677829.81	5	2535565.96	218.25	<0.0001
pH	11310631.13	1	11310631.13	973.57	<0.0001
Zinc	347360.98	1	347360.98	29.90	<0.0001
Potasio	84466.18	1	84466.18	7.27	0.0103
Salinidad	378485.90	1	378485.90	32.58	<0.0001
Salinidad^2	556885.62	1	556885.62	47.93	<0.0001
Error	453088.40	39	11617.65		
Total	13130918.21	44			

Tabla de análisis de la varianza SC Tipo III

FV	SC	gl	CM	F	Valor p
Modelo	12677829.81	5	2535565.96	218.25	<0.0001
pH	1050548.66	1	1050548.66	90.43	<0.0001
Zinc	966936.57	1	966936.57	83.23	<0.0001
Potasio	106199.51	1	106199.51	9.14	0.0044
Salinidad	935371.52	2	467685.76	40.26	<0.0001
Error	453088.40	39	11617.65		
Total	13130918.21	44			

Se puede ver que ahora, además de pH, Salinidad y Zinc, la regresora Potasio presentó un valor $p<0.05$, es decir presenta relación lineal significativa. Además, con la incorporación del término cuadrático para Salinidad, el *Cp* de Mallows para Potasio aumentó de 5.95 (Tabla 40) a 13.94 (Tabla 41), lo que sugiere que esta regresora tiene importancia predictiva en el modelo que incorporó el término cuadrático para salinidad. Para verificar la adecuación del último modelo, se realizaron los gráficos de residuos estudentizados (RE) versus cada una de la regresoras (Figura 17) y en ellos puede verse que no existen tendencias que sugieran falta de ajuste.



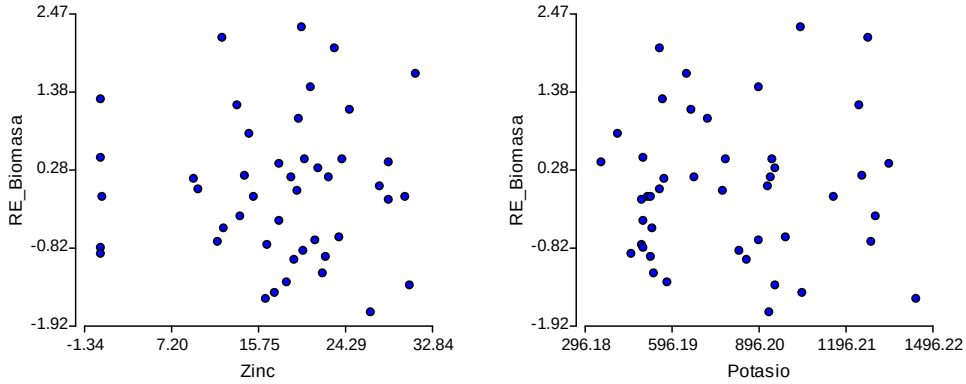


Figura 17: Gráficos de residuos estudentizados de biomasa versus cada una de las variables regresoras. Archivo Salinidad.

Regresión con variables auxiliares (dummy)

Usualmente en la regresión las variables independientes son cuantitativas, pero puede surgir la necesidad de incluir en el modelo variables de clasificación. Las variables de clasificación se incorporan frecuentemente cuando el conjunto de datos a analizar contiene información sobre la relación de interés para subgrupos identificados por una o más variable de clasificación. Una opción en casos donde hay más de un grupo de datos y se quiere analizar la relación funcional entre dos o más variables es ajustar tantos modelos de regresión como grupos existan. Sin embargo ésta no es la técnica más eficiente ya que ningún ajuste utiliza toda la información disponible. Los grados de libertad para el término de error serán más si se ajusta un único modelo a partir de todos los datos, pero este modelo debe indicar la presencia de los grupos para evitar que el efecto grupo interfiera sobre la estimación del modelo de la relación funcional entre la variable dependiente y las regresoras. Esto puede hacerse utilizando variables *auxiliares* (*dummy)*. Para crear las variables auxiliares, se puede usar el generador de variables auxiliares de InfoStat. En el menú DATOS, elegir submenú **Crear variables auxiliares (Dummy).** Para más información ver **Crear variables auxiliares** en el capítulo Manejo de datos.

Ejemplo 25: La tabla del archivo Polímero, presenta los valores de turbidez del medio (Y) y su pH para tres tipos de polímeros A, B, C. El interés se centra en la dependencia de la turbidez del medio respecto al pH del mismo. Al proponer un modelo de regresión lineal simple para explicar la turbidez en función del pH, usando todos los datos y sin indicar que los mismos pertenecen a 3 polímeros (sin indicar la presencia de grupos) se obtienen residuos estudentizados que graficados en función del pH y usando la variable Polímero como criterio de partición, se visualizan de la siguiente manera:

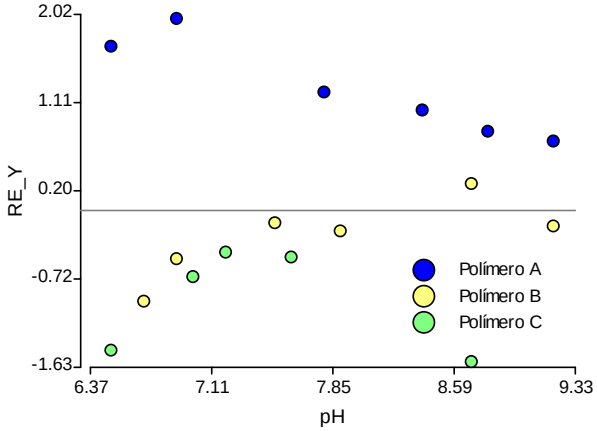


Figura 18: Gráfico residuos estudentizados versus pH para datos del Archivo Polímero.

Colocando una línea de corte en el 0 para los residuos se ve claramente que este no es un patrón deseable para una gráfica de residuos vs. predichos. No existe un patrón aleatorio sino que por el contrario se muestra un comportamiento de los residuos asociado al tipo de polímero. Se desea entonces, proponer un modelo de regresión que considere que las observaciones se encuentran agrupadas o clasificadas debido a la existencia de los diferentes polímeros.

Se incorpora el efecto polímero (variable de clasificación con tres niveles) al modelo a través del uso de variables auxiliares, postulando el modelo:

$$Y = \beta_0+\beta_1 pH+\beta_2 D1+\beta_3 D2+\beta_4 D1*pH+\beta_5 D2*pH+\varepsilon$$

donde D1 y D2 representan dos variables auxiliares y pH es la variable regresora cuantitativa. El número de variables auxiliares D_i a incluir es igual al número de niveles del factor de clasificación que se desea modelar menos 1. Cada variable auxiliar es una variable dicotómica que asume el valor 1 sólo para un nivel del factor de clasificación (en este ejemplo, cada variable auxiliar deberá valer 1 sólo para un tipo de polímero). El conjunto de variables auxiliares asociadas a un factor permiten clasificar las observaciones de acuerdo a los niveles del factor. Observe que D1=1 y D2=0 representan a los datos asociados al polímero A, D1=0 y D2=1 representan a los datos del polímero B y D1=0 y D2=0 a los datos del polímero C. El polímero C es tomado como *nivel de referencia* ya que todas las variables auxiliares incluidas asumen el valor 0 para este polímero (ver Crear variables auxiliares en Manejo de Datos).

La inclusión de variables indicadoras del tipo de polímero permite estimar la diferencia en la turbidez promedio de los polímeros pero no es suficiente para establecer si la relación entre pH y turbidez es diferente para los distintos polímeros. Esa posible diferencia de pendientes puede estudiarse al incluir en el modelo términos que involucren los productos entre las variables auxiliares y la regresora. Luego los coeficientes para las variables generadas por el producto entre una regresora y una variable auxiliar permiten obtener pruebas de

homogeneidad de pendientes. Por ejemplo, si la variable auxiliar D1 vale 1 para el polímero A, la prueba para establecer si la pendiente de la recta para el polímero A es igual o distinta a la correspondiente al polímero de referencia será obtenida a partir del coeficiente D1*pH. Para ajustar el modelo de regresión que incluye las variables auxiliares y las interacciones entre éstas y la o las regresoras, el archivo de datos debe contener a las variables auxiliares, (en este ejemplo, D1 y D2) y a sus productos con la regresora (en este ejemplo, D1*pH y D2*pH). Los productos entre las variables auxiliares y la regresora se obtienen al crear las variables dummies (menú DATOS ⇒ CREAR VARIABLES AUXILIARES), si en el panel **Multiplicar por…** se incluye a la variable regresora, como puede verse en la siguiente tabla:

Tabla 42: Datos con variables auxiliares. Archivo Polímero.

Y	pH	Polímero	Polímero_A	Polímero_B	Polímero_A _pH	Polímero_B_pH
292	6.5	A	1	0	6.5	0
329	6.9	A	1	0	6.9	0
352	7.8	A	1	0	7.8	0
378	8.4	A	1	0	8.4	0
392	8.8	A	1	0	8.8	0
410	9.2	A	1	0	9.2	0
198	6.7	B	0	1	0	6.7
227	6.9	B	0	1	0	6.9
277	7.5	B	0	1	0	7.5
297	7.9	B	0	1	0	7.9
364	8.7	B	0	1	0	8.7
375	9.2	B	0	1	0	9.2
167	6.5	C	0	0	0	0
225	7	C	0	0	0	0
247	7.2	C	0	0	0	0
268	7.6	C	0	0	0	0
288	8.7	C	0	0	0	0
342	9.2	C	0	0	0	0

Para obtener la regresión, elegir Menú ⇒ ESTADÍSTICAS ⇒ submenú REGRESIÓN LINEAL. En la ventana **Análisis de regresión lineal** declarar Y como **Variable dependiente** y a pH, Polímero_A, Polímero_B, Polímero_A_pH y Polímero_B _pH como **Variables regresoras**. El cuadro de **Resultados** mostrará la siguiente salida:

Tabla 43: Análisis de regresión lineal con variables auxiliares. Archivo Polímero.

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP
Y	18	0.97	0.96	556.03

Coeficientes de regresión y estadísticos asociados

Coef	Est.	EE	LI(95%)	LS(95%)	T	p-valor	CpMallows
const	-158.27	48.52	-263.98	-52.57	-3.26	0.0068	
pH	53.82	6.25	40.20	67.45	8.61	<0.0001	73.46
Polímero_A	197.69	68.79	47.80	347.58	2.87	0.0140	12.70
Polímero_B	-108.74	71.05	-263.55	46.07	-1.53	0.1518	7.24
Polímero_A_pH	-13.56	8.74	-32.60	5.48	-1.55	0.1466	7.30
Polímero_B_pH	17.39	9.09	-2.41	37.20	1.91	0.0798	8.46

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	82707.78	5	16541.56	77.76	<0.0001
pH	15759.53	1	15759.53	74.08	<0.0001
Polímero_A	1756.64	1	1756.64	8.26	0.0140
Polímero_B	498.26	1	498.26	2.34	0.1518
Polímero_A_pH	512.47	1	512.47	2.41	0.1466
Polímero_B_pH	778.95	1	778.95	3.66	0.0798
Error	2552.67	12	212.72		
Total	85260.44	17			

La tabla con los coeficientes de regresión brinda la información necesaria para: 1) construir la ecuación de ajuste para cada polímero y 2) realizar pruebas de igualdad de efectos promedios entre los distintos grupos y de homogeneidad de pendiente.

Como cada polímero está representado por una combinación de las variables auxiliares, para lograr los ajustes habrá que tener presente los valores de D1 y de D2 que corresponden a cada polímero.

Para obtener la ecuación de ajuste para el polímero A escriba el modelo completo y reemplace la variable D1 (en el ejemplo Polímero_A) por 1 y la variable D1*pH por pH ya que D1 es 1 para este grupo de datos. Todos los términos del modelo completo que incluyen la variable D2 (en el ejemplo Polímero_B) deberán ser excluidos para obtener la recta de ajuste del polímero A ya que D2 vale 0 para este polímero. Reordenando los términos sin regresora y los términos en la regresora pH se tendrá finalmente la ecuación para el polímero A, como se muestra a continuación:

$$\hat{y} = -158.27 + 53.82\,pH + 197.69D1 - 13.56D1 * pH - 108.74D2 + 17.39D2 * pH$$
$$\hat{y} = -158.27 + 53.82\,pH + 197.69 * 1 - 13.56\,pH$$
$$\hat{y} = 39.42 + 40.26\,pH$$

En forma similar se pueden obtener las ecuaciones para los restantes polímeros:

para el polímero B $\hat{y} = -267.01 + 71.21\,pH$

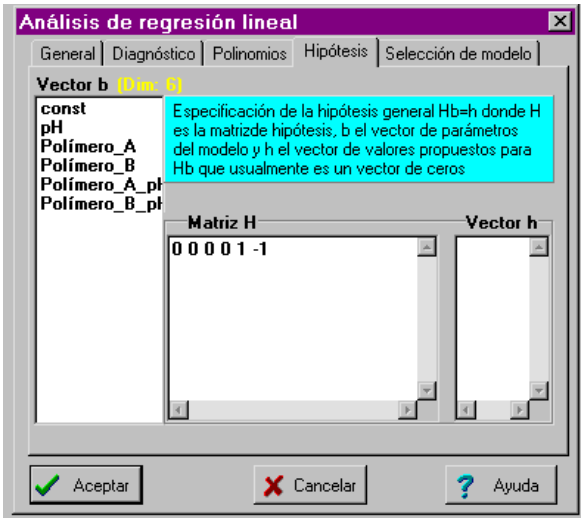
para el polímero C $\hat{y} = -158.27 + 53.82\,pH$

Estas ecuaciones son las mismas que se obtendrían haciendo las regresiones separadamente, usando la variable polímero como criterio de partición, pero los errores estándar de las estimaciones son notablemente menores en el modelo incluyendo las variables auxiliares ya que se trabaja con más grados de libertad para la estimación del error experimental que cuando se realizan los ajustes por separado (uno para cada polímero).

En la tabla del análisis de la varianza los valores *p* para las interacciones Polímero_A_pH y Polímero_B_pH indican los resultados de la comparación de las pendientes del polímero A con C y B con C, respectivamente. Para un nivel de significación del 10% los polímeros B y C tienen pendientes diferentes, mientras que no hay diferencias entre las pendientes de A y C. Para saber si hay diferencias entre la turbidez promedio bajo cada condición, deben observarse los valores *p* para Polímero_A y Polímero_B. En el primer caso $p<0.01$ indica

que hay diferencias entre la turbidez promedio bajo los polímeros A y C. Lo contrario se observa al comparar B y C. La linealidad de la relación es altamente significativa ya que el valor *p* para la variable pH es menor a 0.01.

Cuando existe interés en comparaciones que no son provistas por el análisis de la varianza del modelo completo, pueden realizarse contrastes. Así, por ejemplo, la comparación de las pendientes de los polímeros A y B puede realizarse utilizando los coeficientes 1 y -1 para los términos Polímero_A_pH y Polímero_B_pH, y 0 para los restantes términos del modelo. Estos coeficientes se ingresan en la matriz H de la solapa **Hipótesis** de la ventana **Análisis de regresión lineal**. InfoStat muestra automáticamente los términos del modelo para que el usuario indique en la subventana **Matriz H** los coeficientes del contraste que permite probar la hipótesis de interés.



En la subventana **Vector h** se deberá ingresar el valor supuesto para el contraste. Si se usa el valor cero se estará probando la hipótesis de igualdad (diferencia cero) entre los parámetros del modelo con coeficientes 1 y –1. Otro valor podría ser señalado en esta subventana (**Vector h**) si se supone que la diferencia entre los parámetros seleccionados es de una magnitud determinada (distinta de cero). En la subventana **Matriz H** pueden escribirse más de una línea de contraste, para cada fila de la matriz InfoStat reportará una tabla de análisis de varianza conteniendo la suma de cuadrados para el contraste y el correspondiente valor *p* para la hipótesis que se prueba con dicho contraste. Si se desea asignarle un nombre al contraste, poner en ventana **Matriz H** el nombre y seguido de ":" los coeficientes del contraste.

El contraste planteado en la solapa Hipótesis permite probar si existen diferencias de pendientes entre el polímero A y B. Para el ejemplo las diferencias de pendientes son significativas (*p*=0.0049).

Tabla 44: Matriz H y vector h de coeficientes para la hipótesis general Hb=h.

Matriz H y vector h de coeficientes para la hipótesis general Hb=h

F.V.	const	pH	Polímero_A	Polímero_B	Polímero_A_pH	Polímero_B_pH	h
Coef.Comb.Lineal1	0	0	0	0	1	-1	0

Test para la hipótesis

F.V.	SC	gl	CM	F	p-valor
Hipótesis	2524.22	1	2524.22	11.87	0.0049

En el siguiente gráfico se presenta el ajuste para cada polímero:

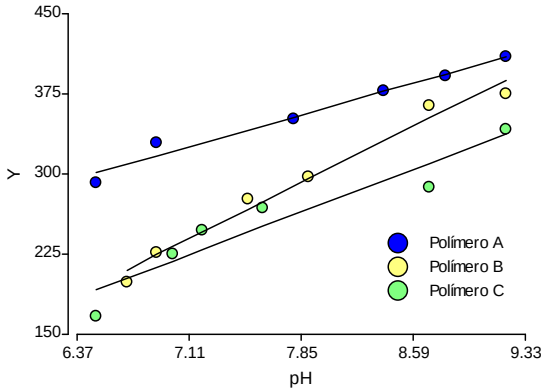


Figura 19: Representación gráfica de las rectas de regresión obtenidas a partir de un modelo con dos variables auxiliares.

Análisis de regresión no lineal

El análisis de regresión no lineal implementado en InfoStat permite obtener los estimadores por mínimos cuadrados de los parámetros de un modelo no lineal arbitrario especificado por el usuario. Como en el caso de regresión lineal el primer paso consiste en seleccionar la variable dependiente y las regresoras, y si fuere necesario incluir un criterio de partición. Indicadas las variables se presenta una ventana de diálogo en la que el sistema requiere que el usuario escriba el modelo que relaciona a la variable dependiente con la/las regresoras. Por ejemplo, si Y es la variable dependiente y las regresoras son **x** y **t**, un modelo podría ser:

$$y = \alpha(x)^{-\lambda t}$$

y el programa espera que se escriba **alfa*x*exp(-lambda*t)**. Una vez que se ha escrito el modelo, hay que validar su sintaxis. Esta acción puede realizarse dando <Enter> al final de la expresión propuesta o apretando el botón de verificación. El proceso de verificación establece si el modelo está libre de errores de sintaxis. En tal caso, identifica los parámetros a estimar y les asigna valores iniciales por defecto. Los parámetros identificados y sus valores iniciales aparecen en la parte inferior derecha de la ventana de diálogo. Los valores iniciales pueden modificarse para proponer un conjunto de valores de partida que facilite la convergencia del algoritmo de estimación. Para editar los valores basta hacer doble *click* sobre el valor que se quiere modificar. Esta acción hará aparecer un campo de edición en el que se puede especificar el valor inicial para cada parámetro, aceptar con <Enter> o hacer doble *click* en otro valor inicial a modificar.

La elección de los valores iniciales de los parámetros es un tema crítico para la regresión no lineal y se le debe prestar especial atención. Algunos modelos no convergen si se parte de

valores iniciales lejanos de aquellos que logran la minimización de la suma de cuadrados o el método puede converger a un mínimo local alejado del óptimo general.

Si se acepta el modelo y sus valores iniciales, la siguiente etapa consiste en la estimación propiamente dicha. Debido a la dificultad que imponen los métodos de regresión no lineal a la estimación, InfoStat aborda el problema en dos fases. La primera consiste en buscar una solución aproximada mediante un método *downhill simplex* propuesto por Nelder y Mead (1965) que no requiere la evaluación de derivadas parciales (se minimiza la posibilidad de errores numéricos). Esta fase termina con una solución o cuando se alcanza el número máximo de iteraciones preestablecidas (500 no modificable por el usuario). La segunda fase implementa el método de Levenberg-Marquardt (Press *et al.,* 1986), partiendo de la solución anterior. Este método requiere el cálculo de la matriz *hesiana* necesaria para el cálculo de la matriz de covarianzas de las estimaciones. Esta fase termina cuando la diferencia de la suma de cuadrados entre dos iteraciones sucesivas es menor o igual a 1E-10 o cuando se alcanza el número máximo de iteraciones especificadas por el usuario (por defecto 20).

Si InfoStat encuentra una solución, presentará en la ventana de resultados una tabla incluyendo las estimaciones de los parámetros, sus errores estándar (asintóticos) y una prueba *T* para la hipótesis sobre la nulidad de los respectivos valores paramétricos. Asimismo, InfoStat provee un resumen de la cantidad de datos incluidos en el análisis, una estimación de la varianza del error y el número de iteraciones en que se alcanzó la solución por el método de Levenberg-Marquardt. Si el número de iteraciones coincide con el número máximo de iteraciones especificado (20 por defecto), el resultado mostrado puede no ser correcto ya que el algoritmo no convergió. Se recomienda, en ese caso, repetir el análisis con un número mayor de iteraciones y/o partir de un conjunto diferente de valores iniciales para los parámetros.

Modelos predeterminados

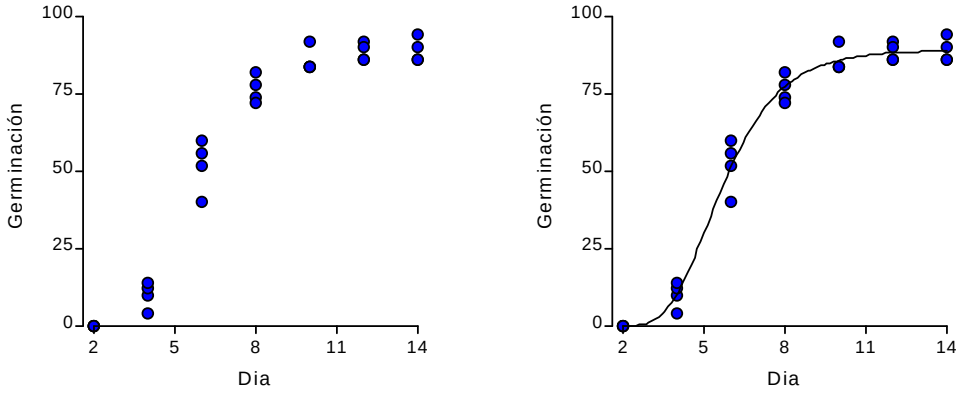
Cuando InfoStat reconoce que hay una única regresora disponible ofrece un conjunto de funciones no lineales comúnmente usadas en modelación. Esta funciones son: Logística, Logística con corrimiento, Gompertz, Gompertz con corrimiento, Exponencial, Monomolecular y Richards. Estas funciones modelan tendencias acumuladas. También están disponibles sus derivadas para modelar velocidades de progreso del fenómeno en estudio (crecimiento, difusión, progreso de enfermedades, etc.). Cuando InfoStat ajusta un modelo para una única regresora provee también una visualización del ajuste, graficando los puntos (x,y) observados, superponiéndoles la función ajustada.

Ejemplo 26: Los datos corresponden al porcentaje de germinación acumulado entre el día 2 y 14 después de la siembra de semillas de un arbusto forrajero sometido a estrés hídrico leve en cuatro ensayos independientes (datos: gentileza Dra. Aiazzi, Facultad de Ciencias Agropecuarias-U.N.C.). El objetivo es modelar la evolución del porcentaje acumulado de germinación en función del tiempo. Los datos se encuentran en el archivo Germinación.

Observación: InfoStat requiere un archivo con dos columnas, una que contenga el día de la observación (variable X) y otra con los valores de germinación (variable Y).

Un gráfico del porcentaje acumulado de germinación versus el tiempo permite visualizar una curva de forma sigmoidea. Existen varios modelos que pueden ser usados para ajustar este tipo de comportamiento. Cuando se modeló utilizando la función de Gompertz se encontró un buen ajuste.

Para realizar esta operación elija Menú ⇒ ESTADÍSTICAS, ⇒ REGRESIÓN NO LINEAL. En la ventana **Análisis de regresión no lineal** asignar como **Variable dependiente** a "germinación" y como **Variable independiente** a "día". En la ventana de diálogo, específica del análisis de regresión no lineal, seleccionar en **Modelos no lineales con solo una regresora** el modelo de **Gompertz**. Finalmente apretar el botón **Aceptar** y esperar que se completen las fases de la estimación. En la ventana **Resultados** se mostrarán los valores de los parámetros ajustados. A continuación se presenta el gráfico de germinación acumulada versus días de la germinación junto con el gráfico de la función ajustada y la salida que se muestra en la ventana **Resultados**.



Valores observados (acumulados) de germinación versus día de observación

Valores observados de germinación (acumulados) y ajuste del modelo de Gompertz versus día de observación

Figura 20: Ajustes no lineales. Archivo Germinación.

Tabla 45: Resultados regresión no lineal. Archivo Germinación.

Modelo Germinación:=alfa*exp(-beta*exp(-gamma*Dia))

Variable	N	CMError	Iteración
Germinación	28	18.54	4

Parámetros	Estimación	E.E.	T	p
ALFA	89.07	1.44	62.04	<0.0001
BETA	32.84	9.44	3.48	0.0019
GAMMA	0.68	0.05	12.54	<0.0001

InfoStat reporta la expresión del modelo ajustado y las estimaciones de cada uno de sus parámetros. En este ejemplo todos los términos del modelo realizan una contribución significativa.

La comparación de modelos alternativos de regresión no lineal se basa en varios criterios. En general se busca que el cuadrado medio del error (CMError) sea lo menor posible, que el número de parámetros del modelo sea lo menor posible (principio de simplicidad), que el error estándar de las estimaciones de los parámetros sea lo más pequeño posible y que los coeficientes estimados no se encuentren altamente correlacionados. Finalmente, el gráfico de dispersión de los residuos versus los valores predichos puede también servir como herramienta de adecuación del modelo.

Observación: Los nombres de las variables que se utilizan como regresoras no deben contener símbolos matemáticos, paréntesis u otros caracteres especiales como acentos, eñes o %.

Análisis de correlación

Bajo este título se encuentran métodos de cálculo de coeficientes de correlación muestral, de coeficientes de correlación parcial y de efectos directos e indirectos en un análisis de sendero (*path analysis*). Todos estos métodos suponen que se tienen dos o más variables aleatorias relevadas sobre cada una de las unidades experimentales u observacionales. El interés es obtener una medida de la magnitud (y dirección) de la asociación o covariación de cada par de variables.

Coeficientes de correlación

Menú ⇒ ANÁLISIS DE CORRELACIÓN ⇒ COEFICIENTES DE CORRELACIÓN. En la ventana **Coeficientes de correlación**, especificar las variables para las cuales desea obtener el coeficiente de correlación. La siguiente ventana permite optar entre los coeficientes de correlación de Pearson o de Spearman (Conover, 1999). Los resultados se presentan como una matriz con las siguientes características: 1) El número de filas es igual al número de columnas e igual al número de variables seleccionadas; 2) Los elementos de la diagonal principal son todos iguales a 1 ya que representan la correlación de una variable con si misma; 3) Por debajo de la diagonal principal y en la posición *i,j* se encuentra el coeficiente de correlación entre la i-ésima y j-ésima variables de la lista; 4) Por encima de la diagonal principal y en la posición *j,i* se encuentra la probabilidad asociada a la prueba de hipótesis de correlación nula entre la j-ésima y i-ésima variables de la lista.

El *coeficiente de correlación de Pearson* es una medida de la magnitud de la asociación lineal entre dos variables que no depende de las unidades de medida de las variables originales. Para las variables j-ésima y k-ésima se define como:

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_j^2 \; S_k^2}} = \frac{\left(\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k) \right) / (n-1)}{\sqrt{\left(\left(\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2 \right) / (n-1) \right) \left(\left(\sum_{i=1}^{n} (x_{ik} - \overline{x}_k)^2 \right) / (n-1) \right)}}$$

donde S_{jk} es la covarianza entre la variable *j* y la variable *k*, S_j^2 y S_k^2 son las varianzas de las variables *j* y *k* respectivamente.

El coeficiente de correlación muestral representa la covarianza de los valores muestrales estandarizados. Asume valores en el intervalo [-1;1] y el signo indica la dirección de la asociación (valores negativos se producen cuando la tendencia promedio indica que si un valor en el par observado es más grande que su media, el otro valor es más pequeño que su media).

El *coeficiente de correlación de Spearman* es una medida no paramétrica de asociación basada en rangos, que puede ser usado para variables discretas o continuas no necesariamente normales. También este coeficiente puede ser usado para medir asociaciones en variables cualitativas ordinales. Para las variables j-ésima y k-ésima se define como:

$$Sr_{jk} = \frac{\sum_{i=1}^{n} R(x_{ij}) \, R(x_{ik}) - n\left(\frac{n+1}{2} \right)^2}{\sqrt{\left(\sum_{i=1}^{n} R(x_{ij})^2 - n\left(\frac{n+1}{2} \right)^2 \right) \left(\sum_{i=1}^{n} R(x_{ik})^2 - n\left(\frac{n+1}{2} \right)^2 \right)}}$$

donde $R(x_{ij})$ es el rango correspondiente a la i-ésima observación de la variable *j* y $R(x_{ik})$ es el rango correspondiente a la i-ésima observación de la variable *k*, con *i*=1,...,*n*.

Coeficientes de correlación parcial

Menú ⇒ ANÁLISIS DE CORRELACIÓN ⇒ CORRELACIONES PARCIALES, permite obtener **Correlaciones parciales** entre dos o más variables después de ajustar por los efectos de una o mas variables adicionales (variables fijadas). Las variable "fijadas" pueden ser variables continuas o de clasificación.

La correlación parcial entre dos variables Y_1 e Y_2 ajustando por X puede interpretarse como el coeficiente de correlación entre Y_2 y los residuos de la regresión de Y_1 sobre X.

Al invocar correlaciones parciales en InfoStat, usando el selector de variables, se deberá indicar las variables para las cuales se desea obtener el coeficiente de correlación parcial (**Variables Y**) y la o las variables por las cuales se deben ajustar las correlaciones (**Variables fijadas**). Al **Aceptar** se obtendrá una matriz con las siguientes características: 1) El número de filas es igual al número de columnas e igual al número de variables Y seleccionadas; 2) Los elementos de la diagonal principal son todos iguales a 1 ya que representan la correlación de una variable con si misma; 3) Por debajo de la diagonal

principal y en la posición *i,j* se encuentra el coeficiente de correlación parcial entre la i-ésima y j-ésima variable Y de la lista ajustada por las variables X fijadas por el usuario.

Coeficientes de sendero (*path analysis)*

Menú ⇒ ANÁLISIS DE CORRELACIÓN ⇒ ANÁLISIS DE SENDERO (PATH ANALYSIS), permite descomponer la correlación entre dos variables (X e Y) en una suma del efecto directo de X sobre Y y los efectos indirectos de X sobre Y vía otras variables independientes del sistema de correlaciones.

Es bien conocido que las correlaciones observadas entre dos variables no pueden ser usadas para establecer relaciones causales. Cuando una variable precede en tiempo a otra y/o se puede postular la existencia de relación causal (y se supone que esta es lineal) se pueden utilizar modelos lineales para expresar dicha relación. El objetivo del *análisis de sendero* o *análisis de ecuaciones estructurales* es proveer posibles explicaciones causales de las correlaciones observadas entre una variable respuesta (dependiente) y una serie de variables predictoras o causales (variables exógenas o independientes).

La noción de una relación causal entre la variable dependiente y una variable independiente requiere que se haya eliminado el efecto de todas las otras variables independientes o causales reconocidas en el sistema. En un modelo de regresión lineal, $Y = \beta_0 + \beta_1 X + \varepsilon$, el último término (término de error aleatorio) representa el efecto colectivo de todas las variables no medidas y que podrían influenciar las variables en estudio. La regresión, escrita en su forma estandarizada, es:

$$\frac{Y-\mu_y}{\sqrt{\sigma_Y^2}} = \beta_1 \sqrt{\frac{\sigma_X^2}{\sigma_Y^2}} \left(\frac{X-\mu_X}{\sqrt{\sigma_X^2}} \right) + \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_Y^2}} \frac{\varepsilon}{\sqrt{\sigma_\varepsilon^2}}$$

que puede expresarse también como:

$$Z_y = p_{yx} Z_X + p_{y\varepsilon} Z_\varepsilon$$

Los parámetros *p* en el modelo estandarizado son conocidos con el nombre de coeficientes de sendero. El modelo causal anterior implica que la correlación entre *X* e *Y* es p_{yx} y que el modelo es auto contenido o completamente determinado dado que las contribuciones de los dos términos del modelo a la varianza de Z_y suman 1, $Var(Z_y) = p_{yx}^2 + p_{y\varepsilon}^2 = 1$.

En el análisis de sendero se pretende construir modelos de causa-efecto entre las variables a través de la partición de la correlación entre dos variables como la suma de dos tipos de efectos. Estos son efectos directos de una variable sobre otra (senderos simples) y efectos indirectos de una variable sobre otra vía una o más variables exógenas (senderos compuestos). Si se considera una nueva variable en el sistema anterior, digamos la variable *U*, y suponemos que existe un sistema con relaciones lineales que pueden ser representadas

por el modelo lineal $Y = \beta_0 + \beta_1 X + \beta_2 U + \varepsilon$, el análisis de sendero nos brindará información sobre los efectos directos de *X* y *U* sobre *Y* (senderos simples en el diagrama del sistema) y además efectos indirectos de *X* sobre *Y* a través de *U* y de *U* sobre *Y* a través de *X*. El efecto indirecto de una variable *X* sobre *Y* vía otra variable *U* se define como $p_u r_{x,u}$, donde los coeficientes *p* corresponden a los coeficientes estandarizados de la regresión múltiple de *Y* sobre *X* y *U* y $r_{x,u}$ es el coeficiente de correlación simple entre *X* y *U*. Luego, el análisis de sendero de este sistema involucrando dos variables causales realiza la siguiente partición de la correlación observada entre *Y* y *X* y de la correlación entre *Y* y *U* (sin considerar los términos de error).

$$r_{y,x} = p_{y,x} + p_{y,u} r_{x,u}$$

$$r_{y,u} = p_{y,x} r_{x,u} + p_{y,u}$$

Dada una muestra, es posible obtener valores para todos los coeficientes de correlación involucrados en este sistema de ecuaciones, el número de incógnitas es siempre igual al número de ecuaciones. Resolviendo el sistema se tienen los estimadores de los efectos directos.

En InfoStat los resultados del análisis de sendero son presentados en tablas donde se muestran todos los efectos directos e indirectos del sistema en estudio. Los coeficientes ayudan a determinar la importancia relativa de los mismos. Las conclusiones del análisis dependerán de la relación lineal supuesta, por ello es oportuno verificar que la correlación entre la variable de salida del sistema y el término de error es baja, implicando que no existen factores causales importantes que no hayan sido incorporados al modelo.

Ejemplo 27: En un experimento sobre crecimiento de una maleza se utilizan 20 unidades experimentales consistentes en bandejas sembradas con 40 semillas al comienzo de la experiencia. Se registra el número de semillas germinadas y al cabo de un cierto tiempo en todas se obtiene un indicador del área foliar y la biomasa total. Se pretende estudiar las correlaciones de biomasa con área foliar y número de semillas germinadas, en un sistema donde la biomasa es considerada como variable dependiente. Los datos se encuentran en el archivo Sendero.

Menú ⇒ ANÁLISIS DE CORRELACION ⇒ ANÁLISIS DE SENDEROS (PATH ANALYSIS), la ventana **Coeficientes de sendero (path analysis)** permite seleccionar las variables dependientes y las variables del sistema. Indicar a "Biomasa" como **variable dependiente** y a "SemGerm" y "AreaFoliar" como **variables independientes**. Al **Aceptar** en la ventana de resultados aparecerá la siguiente información.

Tabla 46: Resultados análisis de sendero. Archivo Sendero.

Coeficientes de Sendero (Path Analysis)

Variable dependiente: Biomasa

Efecto	Via	Coeficientes	valor p
SemGerm	Directa	0.78	
SemGerm	AreaFoliar	-0.02	
r total		0.76	<0.0001
AreaFoliar	Directa	0.03	
AreaFoliar	SemGerm	-0.52	
r total		-0.49	0.0272

La correlación entre biomasa y área foliar es significativa (r=-0.49, p=0.0272) y está casi completamente determinada (-0.52) por la correlación entre biomasa y semillas germinadas.

Correlación entre matrices de distancia

La correspondencia entre dos matrices de distancia (o similitud). La via mas simple es considerar a los elementos de las dos matrices de distancia que son informativos , i.e los $n(n\text{-}1)/2$ elementos distintos fuera de la diagonal y obtener un estimador de la correlacion elemento a elemento, como es el coeficiente de correlacion lineal de Pearson.

Usualmente se acompañan el coeficiente de correlacion con un diagrama de dispersión entre los elementos de una y otra matriz, con el objeto de poner en evidencia pares de datos anómalos en el patron de correlacion. Una prueba mas adecuada para evaluar la correlacion entre dos matrices es la Prueba basada en el estadisitico Z de Mantel, cuya significancia se obtiene por permutación (Mantel, 1967), ya que los pares de datos con los que se calcula el estadistico de correlacion no son realmente independientes. La prueba de Mantel (1967) tiene en cuenta las autocorrelaciones de los elementos de una matriz de distancia. El estadistico de la prueba es:

$$Z = \sum_{i<j}^{n} x_{ij} y_{ij}$$

donde X_{ij} e Y_{ij} son los elementos i,j (elementos fuera de la diagonal principal) de las matrices X e Y, respectivamente. El estadisticazo es la suma de los productos cruzados de esos elemntos, ya que esta es la única cantidad que realmente es sensible a las permutaciones que se realizaran para evaluar la significancia de la asociacion. Si las dos matrices muestran relación de semejanza, entonces Z debería ser grande en comparación al valor de Z que en promedio se esperaria si se utilizan matrices no correlacionadas. El valor observado de Z se posiciona sobre la distribución de Z bajo la hipótesis de que las matrices no estan correlacionadas la cual es obtenida por permutación y se calcula la probabilidad de valores de Z mayores o iguales al observado. Si es probabilidad (valor-p) es menor al nivel de significancia de la prueba, e.g. alfa 0.05, entonces se rechaza la hipótesis de falta de asociación y se concluye que existe cierta correspondencia entre ambas matrices.

La distribución de Z bajo la hipótesis nula, es obtenida mediante la comparación de una matriz, digamos X, con todas las posibles matrices Y en las cuales el orden de los objetos (o variables) ha sido permutado. Luego, el procedimiento de la prueba de Mantel aplicado a dos matrices consistirá entonces en calcular las cantidades de Z_{XY} desde las matrices originales, permutando las filas y columnas de una matriz mientras la otra permanece constante, recalculando cada vez la cantidad de Z^*_{XY}, y comparando este valor con el valor de Z_{XY} original.

Como la permutación del orden de los objetos no tiene efectos sobre la media y la varianza de la matriz Y, es importante notar que el coeficiente de correlación producto-momento de Pearson se encuentra monotónicamente relacionado a Z. Así que, de detectar correlación significativa con la prueba de mantel, es posible reportar el coeficiente de correlación de Pearson como medida de la magnitud de la asociación. Este último tiene la ventaja de estar expresado en unidades más familiar debido a la estandarización y por lo tanto es más fácil de interpretar. InfoStat reporta el estadístico en esta escala.

Datos Categorizados

Tablas de contingencia

Menú ESTADÍSTICAS ⇒ DATOS CATEGORIZADOS ⇒ TABLAS DE CONTINGENCIA, permite construir tablas de clasificación cruzada según diversos criterios de clasificación. Agresti (1990) presenta excelentes tratados sobre análisis de datos categorizados en el que extensamente se cubre el tópico de modelización y análisis de tablas de contingencia. Parte de la terminología comúnmente utilizada en el análisis de dichas tablas se presenta a continuación.

Las tablas de contingencia (formas tabulares de presentar datos categorizados) son útiles para el análisis simultáneo de dos o más variables categorizadas. Una variable categorizada es aquella en la cual la escala de medida consiste en un conjunto de categorías, por ejemplo la variable tipo de vivienda puede ser categorizada de acuerdo a las siguientes dos categorías "rural" y "urbana". Para analizar e interpretar apropiadamente tablas de contingencia es necesario tener en cuenta la escala de medida de las variables involucradas y el tipo de estudio (aleatorización) usado para obtener los datos. Comúnmente, las hipótesis de interés en tablas de contingencia se refieren a la asociación entre las variables que definen las filas y las columnas de la tabla.

Las variables categorizadas con niveles que no tienen un ordenamiento natural se denominan *nominales* (por ejemplo, afiliación política con categorías "liberal" y "conservador"). Un caso particular es aquel de las variables binarias las cuales involucran 2 categorías de variables nominales, por ejemplo, "si" y "no", "respuesta" y "no respuesta".

Si los niveles se encuentran ordenados la variable se denomina *ordinal;* por ejemplo, grado de infección categorizada como "leve", "moderada" y "severa". Si bien las categorías pueden ser ordenadas, a diferencia de las variables cuantitativas las distancias absolutas entre categorías son desconocidas. En algunas situaciones las tablas pueden ser construidas con variables medidas en una escala de intervalos, esta escala implica que se conoce la distancia numérica entre dos niveles cualesquiera de la escala (por ejemplo, intervalos de la variable edad).

Las variables que constituyen la tabla pueden ser consideradas como variables de *respuesta* o como variables de *clasificación*. Las primeras, también llamadas variables *dependientes*, son aleatorias y describen lo que fue observado en las unidades muestrales. Las segundas, también llamadas variables *independientes* o *factores*, son fijas por condicionamiento y las combinaciones de sus niveles definen estratos, poblaciones o subpoblaciones a las cuales las unidades muestrales pertenecen. Cuando todas las variables de la tabla son de respuesta generalmente se analiza la asociación entre ellas. Cuando algunas son respuesta y otras de clasificación, en general se estudian los efectos de las variables de clasificación sobre la *distribución* de las variables de respuesta. Si denotamos por *X* a una variable categorizada con *I* categorías o niveles y por *Y* a otra variable con *J* niveles, para clasificar sujetos sobre ambas variables existirán $I \times J$ combinaciones de clasificación.

Los pares (*X*,*Y*) asociados a cada sujeto seleccionados aleatoriamente desde una población tienen una distribución de probabilidad. La distribución se presenta en una tabla con *I* filas y *J* columnas. La probabilidad asociada al evento *IJ*, en general denotada por π_{ij} representa la probabilidad de que la variable *X* asuma la categoría *I* y la variable *Y* asuma la categoría *J*. El conjunto de los valores π_{ij} forman la *distribución conjunta* de ambas variables. El conjunto de los valores π_{i+} (total de las probabilidades conjuntas de la fila *i*) para $i=1, ..., I$, forma la *distribución marginal* de las filas de la tabla. Equivalentemente se puede obtener la distribución marginal de las columnas. Cuando una variable (digamos, *Y*) es considerada como variable respuesta y la otra (digamos, *X*) como variable explicativa, es informativo identificar las distribuciones de probabilidad de la respuesta para cada nivel de *X*, entiéndase la *distribución condicional* de *Y* dado *X*.

La noción de *independencia* es comúnmente utilizada en tablas de contingencia. Dos variables (*X* e *Y*) son estadísticamente independientes si las distribuciones condicionales de *Y* son idénticas para todos los niveles de *X*. Cuando ambas variables son consideradas como variables respuesta es indistinto observar la distribución condicional de *Y* dado *X* o la distribución condicional de *X* dado *Y*. La independencia estadística expresa las probabilidades conjuntas (probabilidad de la celda *ij*) como el producto de las probabilidades marginales, entiéndase probabilidad de la fila *i* por probabilidad de la columna *j* (valor esperado bajo independencia).

Las tablas de contingencia pueden ser utilizadas para visualizar resultados obtenidos de distintos tipos de estudios: 1) *estudios experimentales*, aquellos donde el investigador tiene control sobre el grupo de sujetos; es decir, decide bajo que condiciones va a ser observado cada sujeto. Estos estudios son de tipo prospectivo y en el campo biomédico se conocen

como ensayos clínicos (*clinical trials*); 2) *estudios observacionales*, los cuales pueden ser retrospectivos (caso-control) o prospectivos (*cohortes, cross-sectional o transversales*). En el tipo caso-control se investiga el pasado seleccionando arbitrariamente un grupo de sujetos que tienen la característica en estudio (casos) y otro grupo de sujetos que no la tienen para ser usados como referencia (control). Esta selección arbitraria impide realizar ciertas inferencias sobre *Y*. La distribución marginal de *Y* está determinada por el muestreo y no necesariamente responde a las características de la población. En el tipo cohorte o transversales se parte de una muestra aleatoria de sujetos los cuales son clasificados en una de las celdas *ij* de la tabla, simultáneamente, según corresponda. Los totales marginales son de tal manera aleatorios (no fijados por el experimentador).

Así, el diseño del estudio implica un tipo de muestreo particular el cual deberá ser tenido en cuenta a la hora de interpretar los estadísticos obtenidos de la tabla de contingencia. Típicamente para tablas 2×2, entiéndase $I=2$ $J=2$, se identifican los siguientes muestreos: 1) *muestreo Poisson*, cada celda es una variable Poisson independiente, derivado de estudios transversales donde el muestreo es aleatorio y el número total de individuos (n) no es fijo; 2) *muestreo binomial*, cada fila de la tabla define diferentes grupos y los tamaños muestrales de la fila son fijados por el diseño (existe condicionamiento), comúnmente se necesita analizar las distribuciones condicionales a las filas las cuales se modelan con una distribución binomial para tablas 2×2 (en caso de tablas con $J>2$ se utiliza el modelo multinomial para cada fila); 3) *muestreo multinomial*, los conteos de las celdas son multinomiales, el tamaño muestral total es fijo pero no se fijan los totales de filas ni de columnas; 4) con n y marginales fijos la distribución de valores por celda puede aproximarse a una distribución hipergeométrica.

Ejemplo 28: La siguiente tabla corresponde a la clasificación de los empleados de una empresa, según la sucursal a la que pertenecen y su opinión sobre las oportunidades de ascenso en sus cargos. Los datos se encuentran en el archivo Categorizados.

Tabla 47: Número de empleados según su opinión sobre la oportunidad de ascenso, en tres sucursales. Archivo Categorizados.

Sucursal	Oportunidad de ascenso			
	Baja	Moderada	Alta	Total
A	205	174	138	517
B	199	184	118	501
C	152	167	227	546
Total	556	525	483	1564

Menú ⇒ ESTADÍSTICAS ⇒ DATOS CATEGORIZADOS ⇒ TABLAS DE CONTINGENCIA, permite identificar en la ventana **Tablas de contingencia** las variables que serán utilizadas para conformar los criterios de clasificación de filas y de columnas. Para declarar las variables del Ejemplo 28 se deberán indicar como **Criterios de clasificación** a las columnas "Sucursal" y "OpAscenso". La variable "Frec" debe ingresarse en la subventana **Frecuencias**. Al **Aceptar**, en la solapa **Selección de filas y columnas** se deberá indicar que "Sucursal" define las filas y "OpAscenso" las columnas de la tabla.

Tabla 48: Resultados del análisis de tablas de contingencia a dos vías de clasificación. Archivo Categorizados.

Tablas de contingencia
En filas: Sucursal
En columnas: OpAscenso

Sucursal	Alta	Baja	Mod	Total
A	138	205	174	517
B	118	199	184	501
C	227	152	167	546
Total	483	556	525	1564

Estadistico	Valor	gl	p
Chi Cuadrado Pearson	48.84	4	<0.0001
Chi Cuadrado MV-G2	48.33	4	<0.0001
Coef.Conting.Cramer	0.10		
Coef.Conting.Pearson	0.17		

Como el valor p<0.0001 se concluye que hay una relación entre la sucursal y la opinión emitida.

InfoStat permite analizar tablas a tres vías de clasificación. Opcionalmente se puede indicar una columna del archivo como *criterio de estratificación* de los datos. Suponga que el mismo experimento se lleva a cabo en dos provincias diferentes; luego existirán dos tablas como la anterior cada una construida a partir de los datos de una provincia particular. La variable *provincia* puede ser señalizada como un *criterio de estratificación* del conjunto total de datos. En tal caso InfoStat trabajará con un conjunto de tablas de contingencia (una tabla por estrato) proveyendo estadísticos clásicos para cada estrato y para la tabla marginal (tabla resumen a través de todos los estratos). Si se asigna más de una variable en la ventana **Estratos (opcional)**, los estratos serán definidos a partir del cruzamiento de dichas variables. La hipótesis de asociación en ambos tipos de encuestas (con y sin estratos) estudia la asociación entre las sucursales (tratamientos) y la opinión con respecto a la posibilidad de ascenso (respuesta), sólo que cuando existen estratos esta asociación es deducida después de controlar o ajustar por el efecto del estrato.

Organización de los datos

Dos formatos diferentes de archivos pueden ser usados para construir tablas de contingencia:

Formato expandido: cada caso contiene los valores registrados para las variables X y Y sobre cada unidad de estudio o sujeto. Por ejemplo si se tienen como variables de clasificación: “sexo” (2 niveles) y “fuma” (2 niveles) y se registran 45 datos para cada sexo, habrá un total de 90 datos; la observación registrada para cada individuo conformará un caso. El archivo tendrá dos columnas (“sexo” y “fuma”) y en cada una de las 90 filas “sexo” asumirá el valor F o M y “fuma” el valor Si o No según haya sido la respuesta del individuo en cuestión.

Formato usando frecuencias: si se dispone de los recuentos de cada celda de la tabla se puede utilizar un formato de tres columnas, una de ellas identificando los niveles de las

filas, otra con los niveles de la variables columna y la tercera con los recuentos de cada combinación de niveles filas por columna. Para el ejemplo anterior, la tabla tiene un total de 4 celdas, usando las frecuencias en cada celda, los datos se ingresarían de la siguiente manera:

Las variables (columnas de la tabla) "Sexo" y "Fuma" serán asignadas como **Criterios de clasificación** y la columna **Frecuencias** (ver también archivo *Contingencia*) en **Frecuencias (opcional-sólo una)**. Si la relación entre sexo y el hábito de fumar se realizara en 3 poblaciones diferentes, una cuarta columna identificando la procedencia (población) de cada recuento podría ser listada como **Estrato**.

Caso	Sexo	Fuma	Frecuencias
1	F	SI	30.00
2	F	NO	15.00
3	M	SI	35.00
4	M	NO	10.00
5			

La segunda ventana contiene dos solapas: **Selección de filas y columnas** y **Opciones**. En la primera aparecerán listadas las variables del archivo seleccionadas como variables de clasificación para que el usuario indique cuál variable debe ser usada para construir las filas de la tabla y cuál para las columnas de la misma. Pueden especificarse una o más variables para filas y columnas.

En caso de tener más de una variable asignada a filas (columnas), la tabla tendrá tantas filas (columnas) como combinación de niveles tengan las variables seleccionadas. Por ejemplo, si una variable es "sexo" y la otra es "grupo sanguíneo" y ambas variables son asignadas a filas, entonces la tabla tendrá a lo sumo 8 filas (2 sexos)×(4 grupos sanguíneos). En la misma ventana si se activa la opción **Todas las tablas de a pares**, InfoStat ignora todas las combinaciones de niveles de las variables declaradas en Filas y Columnas y calcula las tablas de contingencia para todos los pares de las variables listadas en filas y columnas. Si no se asignan variables a las filas, la tabla tendrá sólo una fila para incluir la información de las columnas y viceversa. Si se mantiene la opción **Presentación en orden alfabético** (activada por defecto), la tabla de contingencia se construirá presentando los valores de las variables en filas y columnas en orden alfanumérico. Al desactivar esta opción las tablas serán producidas con el orden de las filas y columnas respondiendo a la forma en que fueron ingresados los datos en el archivo. La opción debe ser considerada cuidadosamente cuando se trabaja con variables respuesta ordinales cuyas modalidades en orden alfanumérico no se correspondan con el ordenamiento natural de las mismas, por ejemplo, la variable nivel de infección categorizada como baja, moderada y alta donde el orden alfabético (alta, baja, moderada) no se corresponde con la ordinalidad de las categorías.

En la solapa **Opciones** se puede elegir la información que se desea visualizar en la tabla de contingencia y los estadísticos a reportar. Es posible obtener tablas conteniendo: **Frecuencias absolutas**, **Frecuencias relativas por filas**, **por columnas** y **en relación al total**. También se puede solicitar: **Frecuencias esperadas bajo independencia**, **Desviaciones respecto de lo esperado bajo independencia**, **Desviaciones respecto de lo esperado bajo indep. estandarizadas**, la cual se construye a partir de la tabla anterior

dividiendo por la raíz cuadrada del valor esperado, y **Residuos ajustados**. Las frecuencias relativas son por defecto reportadas como valores en el intervalo [0,1], si se desea visualizar esta información en forma de porcentaje se debe activar la opción **Frecuencias relativas como porcentajes**.

Para tablas de cualquier dimensión se puede solicitar estadísticos para pruebas de hipótesis aproximadas basadas en la distribución Chi cuadrado activando la opción **Chi cuadrado**. En tal caso InfoStat reportará los valores de los estadísticos Chi cuadrado de Pearson, Chi cuadrado máximo verosímil o estadístico G^2 (**Chi cuadrado MV-G2**), el coeficiente de contingencia de Cramer (**coef. conting. Cramer**), el coeficiente de contingencia de Pearson (**coef. conting. Pearson**) y los valores *p* de las pruebas de hipótesis respectivas. Todos los estadísticos miden tipos generales de asociación.

En tablas $I \times J$ con muestreo multinomial, la hipótesis de independencia estadística implica que la frecuencia esperadas para la celda *ij* corresponde al producto de las frecuencias marginales de la fila *i* y la columna *j*, para probar esta hipótesis InfoStat provee del estadístico Chi cuadrado de Pearson, cuya expresión es:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

donde n_{ij} representa al recuento muestral de la celda *ij* y $\hat{m}_{ij}$ es el estimador de la frecuencia absoluta esperada obtenida como $\hat{m}_{ij} = np_{i.}p_{.j}$ con $p_{i.}$ y $p_{j.}$ las frecuencias relativas de las filas *i* y columna *j*, respectivamente. El estadístico se distribuye (bajo la hipótesis nula) como una Chi cuadrado con (*i*-1)(*j*-1) grados de libertad. Valores *p* altos (mayor al nivel de significación nominal de la prueba) implican que no existe suficiente evidencia muestral para rechazar la hipótesis de independencia entre la variable fila y la variable columna. Si el valor *p* conduce al rechazo de la hipótesis nula de independencia entonces se concluirá que existe asociación entre ambas variables.

El estadístico Chi cuadrado con *v* grados de libertad puede ser particionado en *v* componentes Chi cuadrado independientes con un grado de libertad cada una. Este resultado permite amalgamar columnas de la tabla para probar independencia respecto otra columna con un estadístico Chi cuadrado con (*i*-1) grados de libertad.

El tamaño de muestra requerido para el estadístico Chi cuadrado de Pearson establece que todos los valores esperados bajo la hipótesis de independencia debieran ser mayores o iguales a 5. La aplicación de este criterio a otros estadísticos para medir asociación que toman ventaja de la naturaleza de los datos puede resultar muy conservador (Agresti, 1990).

InfoStat también provee de la prueba basada en el cociente de máxima verosimilitud (**Chi Cuadrado MV-G2**) para la hipótesis discutida anteriormente. El estadístico, denotado como G^2, es:

$$G^2 = 2\sum_i \sum_j n_{ij} \log(n_{ij} / \hat{m}_{ij})$$

El estadístico bajo la hipótesis nula también se distribuye para muestras grandes como una Chi cuadrado con (i-1)(j-1) grados de libertad. Luego Chi cuadrado y G^2 son asintóticamente equivalentes. Los resultados asintóticos obtenidos asumiendo un muestreo multinomial también se mantienen para otros tipos de muestreo (Agresti, 1990). Ambos estadísticos son invariantes a permutaciones del orden de filas y/o columnas (tratan a la variable categorizada como nominal).

Otras medidas de asociación provistas por InfoStat son los coeficientes basados en el estadístico Chi cuadrado (valores índices que resumen asociación). Estos son el **Coeficiente de contingencia de Cramer** y el **Coeficiente de contingencia de Pearson** que se calculan de la siguiente manera:

Coeficientes de contingencia de Cramer : $V^2 = X^2 /(n \min(i-1, j-1))$

Coeficientes de contingencia de Pearson : $PCC = (\chi^2 /(\chi^2 + n))^{1/2}$.

Los valores de ambos coeficientes se encuentran entre 0 y 1, valores cercanos a cero implican independencia de los valores fila y columna de la tabla. En tablas 2×2, además de requerir los estadísticos mencionados anteriormente es posible solicitar, para muestreos aleatorios, la prueba exacta de Fisher (**Irwin-Fisher**) y las siguientes medidas de asociación: cocientes de chance (**odds ratios**), **Riesgo relativo** y **coeficiente Phi**. Para el caso particular de muestras no independientes se puede obtener la prueba de **McNemar**.

La prueba exacta de Fisher permite probar la hipótesis de independencia sin necesidad de trabajar con aproximaciones asintóticas. La distribución exacta del estadístico bajo la hipótesis nula se basa en la distribución de los recuentos condicionando sobre las frecuencias marginales. La distribución de las frecuencias observadas es la distribución hipergeométrica y no depende de ningún parámetro desconocido. La distribución exacta puede ser expresada en términos de los conteos en la celda 1,1 (n_{11}). El rango de posibles valores de n_{11} con marginales fijos es conocido y luego es posible construir todas las tablas posibles bajo la hipótesis de independencia y calcular en forma exacta la probabilidad de obtener valores del estadístico mayores al observado a partir de la lista de dichas tablas. InfoStat provee los valores *p* exactos para pruebas de independencia bilateral y unilaterales (derecha e izquierda) en tablas 2×2.

Los cocientes de chance (*odds ratios*) son medidas de asociación frecuentemente usadas. El cociente de chance en una tabla 2×2 se define como el cociente de los productos de recuentos que se encuentran en diagonal en la tabla, por eso también se lo suele llamar cociente de productos cruzados. La estimación muestral del cociente de chance es,

$$\hat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

Los cocientes de chance muestrales pueden no estar definidos en casos donde las dos entradas en una fila o columna son cero. Como este evento tiene probabilidad positiva, el valor esperado y la varianza del estimador no existen (Agresti, 1990). Usando transformación logarítmica y expresando el estimador mediante una estructura aditiva más que multiplicativa, se demuestra para los tipos de muestreo descriptos, que el estimador de cociente de chances se distribuye asintóticamente normal y su error estándar asintótico es:

$$\hat{\sigma}(\log \hat{\theta}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)^{1/2}$$

InfoStat utiliza esta expresión reemplazando n_{ij} por (n_{ij}+0.5) debido a que dicha sustitución mejora el estimador (Agresti, 1990). Además provee, usando la aproximación normal, un intervalo de confianza (con coeficiente de confianza 95%) para el cociente de chance calculado, usando el grupo 1 en el numerador del mismo (1/2) y su recíproco, es decir el cociente de chance de éxitos, en el grupo 2 respecto al 1 (2/1).

Por ejemplo, un cociente de chance de 3 con un intervalo de confianza que no incluya el 1 implican que existe asociación significativa entre la variable fila y la variable columna. El valor 3 significa que el grupo de casos en la fila 1 tienen triple chance de registrar un "éxito" en la variable respuesta que aquellos del grupo 2. Esto equivale a decir que los casos en el grupo 2 tienen un tercio de la chance de registrar un "éxito" en la variable respuesta que los casos en el grupo 1.

Para comparar la distribución condicional de una columna de la variable respuesta dentro de cada fila, se provee de los estadísticos conocidos como *riesgo relativo*. Si los recuentos en cada fila de la tabla se distribuyen como una binomial (las dos filas son binomiales independientes), la diferencia de proporciones en una columna entre las dos filas puede ser evaluada a través de simples intervalos de confianza provenientes de la aproximación normal (ver Diferencia de proporciones en menú Estadísticas, submenú Inferencia basada en dos muestras). El riesgo relativo muestral (RR) es obtenido como:

$$RR = \frac{p_{1/1}}{p_{1/2}}$$

donde $p_{i/j}$ es el estimador de la probabilidad de obtener la respuesta i dado que el caso pertenece a la fila j. El riesgo relativo siempre es mayor o igual a cero, valores iguales a 1 sugieren que las probabilidades del tipo de respuesta en cuestión son las mismas para ambas filas. El RR puede tener valores altos cuando ambas probabilidades se encuentren cercanas a 0 o 1, mientras que para tales casos la diferencia de proporciones podría ser muy pequeña. Por ejemplo, si la probabilidad de éxito en la fila 1 es 0.01 y en la fila 2 es 0.001, la diferencia de proporciones será de 0.009 mientras que el RR es 10.

El coeficiente Phi también mide asociación, se basa en el estadístico Chi cuadrado y se calcula sólo para tablas 2×2 (Conover, 1999). El coeficiente de contingencia de Cramer en tablas 2×2 corresponde a la medida de asociación conocida como *Phi cuadrado* y también

es equivalente al *tau de Goodman y Kruskal* (Agresti, 1990) (medida de reducción proporcional en variación).

La prueba de McNemar es utilizada para evaluar la significancia de cambios en la categoría de respuestas bivariadas (observaciones apareadas para cada individuo). Se asume que se dispone de *n* pares mutuamente independientes de variables aleatorias (X_i,Y_i) categorizadas dicotómicas. Luego, los posibles valores de la observación i-ésima son (0,0), (0,1), (1,0) y (1,1) que habitualmente se presentan en como las celdas de una tabla de contingencia de 2×2. El estadístico de McNemar tiene la forma:

$$T_1 = \frac{(b-c)^2}{b+c}$$

donde *a*, *b*, *c* y *d* es el número de recuentos u observaciones de la muestra clasificadas en las celdas (0,0), (0,1), (1,0) y (1,1) respectivamente. Este estadístico no depende de *a* y *d* ya que éstos representan el número de casos donde ambas variables asumen el mismo valor y son descartados del análisis porque se pretende estudiar cambios en las modalidades de las variables. Como ejemplo, Conover (1999) presenta una situación donde se realiza una encuesta sobre intención de voto para dos partidos políticos, antes y después de un debate televisivo entre los candidatos de ambos partidos. Interesa evaluar si después del debate hubo cambios significativos de la intención de votos. Si con el valor cero de X se representa la intención de votos para el candidato A antes del debate y con el valor cero de Y la intención de votos para el candidato A después del debate, las hipótesis que se prueban son:

H_0: $P(X_i=0, Y_i=1)=P(X_i=1, Y_i=0)$

H_1: $P(X_i=0, Y_i=1)\neq P(X_i=1, Y_i=0)$ para i=1, …,*n*

InfoStat utiliza la aproximación asintótica de la distribución del estadístico T_1, la cual es una Chi cuadrado con un grado de libertad cuando $b+c$ es grande y cuando $b+c$ es menor a 200 deriva los valores de probabilidad a partir de la distribución exacta del estadístico $T_2=b$ cuya distribución es Binomial con parámetros $(b+c, 0.5)$.

Tablas de contingencia a tres vías de clasificación

En estudios de asociación entre una variable respuesta Y y otra variable explicativa X donde existen una o más variables de control (Z) que definen estratos, es necesario usar la información sobre estratificación para estudiar la asociación entre X e Y sin confundimiento debido a la presencia de los estratos. Una forma de eliminar el efecto de Z es trabajar las asociaciones entre X e Y para cada nivel de Z (tablas parciales por estrato). Otra es trabajar con todos los datos pero descontando el efecto de Z.

InfoStat permite generalizar el análisis de tablas a dos vías para tablas a tres o más vías mediante: 1) la construcción de todas las tablas parciales, es decir, aquellas correspondientes a cada nivel de Z. Para ello activar **Mostrar tablas parciales**, que efectivamente remueve el efecto de Z, 2) la construcción de la *tabla marginal*, es decir, aquella donde para cada celda se presenta la suma de los conteos de las celdas de las tablas parciales (suma sobre Z),

ignora el efecto de Z, 3) prueba de Cochran-Mantel-Haenzel (Agresti, 1990) para la asociación entre X e Y, controlando por Z y en casos de tablas 2×2 el cociente de chance de Mantel-Haenzel. Estas pruebas y los estadísticos asociados contrastan las hipótesis sobre la relación entre X e Y utilizando toda la información de las tablas generadas por los niveles de Z. El estadístico de la prueba de Cochran-Mantel-Haenzel (CMH), que se distribuye como una Chi cuadrado con un grado de libertad, combina la información de todas las tablas parciales y esta dado por:

$$CMH = \frac{\left[\sum_k \left(n_{11k} - \mu_{11k}\right)\right]^2}{\sum_k Var\left(n_{11k}\right)}$$

donde la sumatoria en *k* es la suma a través de todas la tablas parciales, n_{11k} es el recuento en la celda 1,1 de la k-ésima tabla, μ_{11k} y $Var(n_{11k})$ son la esperanza y la varianza del recuento en la celda 1.1 respectivamente. El cociente de chance de Mantel-Haenzel (que controla por Z) esta dado por:

$$\hat{\theta}_{MH} = \frac{\sum_k \left(n_{11k} n_{22k} / n_{++}\right)}{\sum_k \left(n_{12k} n_{21k} / n_{++}\right)}$$

donde la sumatoria en *k* es la suma a través de todas la tablas parciales, n_{ijk} es el recuento en la celda i,j de la k-ésima tabla y n_{++} es el total de observaciones en la misma tabla.

Ejemplo 29: En un estudio transversal donde intervinieron 676 estudiantes, se tuvo en cuenta la condición de aprobación (1) o no (2) del examen de ingreso ("AprobExamen") y el tipo de preparación para el examen: el alumno se preparó solo (2) o en una academia (1) ("Preparación"). Como los alumnos provenían de dos facultades, se desea estudiar la relación entre aprobación y tipo de preparación controlando por la unidad académica de origen ("Facultad"). Los datos se encuentran en el archivo IngresoUniv.

Menú ESTADÍSTICAS ⇒ DATOS CATEGORIZADOS ⇒ TABLAS DE CONTINGENCIA, permite obtener la tabla de contingencia 2×2 asociada a este problema y los estadísticos de asociación correspondientes. Se seleccionó "AprobExamen" y "Preparación" como **Criterios de clasificación**; "Facultad" como **Estrato** y "Recuentos" como **Frecuencias**. En la siguiente ventana se designó "AprobExamen" para las columnas y "Preparación" para las filas, con lo cual se obtiene la salida de la Tabla 49.

Tabla 49: Resultados análisis de tablas de contingencia a tres vías de clasificación. Archivo IngresoUniv.

Tablas de contingencia
Frecuencias: Recuentos
Criterio de Estratificación: Facultad

Tabla Marginal
Frecuencias absolutas
En columnas: AprobExamen

Preparación	1	2	Total
1	54	16	70
2	434	172	606
Total	488	188	676

Estadísticos para la tabla marginal

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	0.95	1	0.3286
Chi Cuadrado MV-G2	0.99	1	0.3200
Irwin-Fisher bilateal..	0.06		0.3982
Coef.Conting.Cramer	0.03		
Coef.Conting.Pearson	0.04		
Coeficiente Phi	0.04		

Cocientes de chance (odds ratio) y riesgos relativos

Estadístico	Estim	LI 95%	LS 95%
Odds Ratio 1/2	1.34	0.75	2.38
Odds Ratio 2/1	0.75	0.42	1.33

Estadísticos corregidos por efecto de estrato
Prueba de Cochran-Mantel-Haenszel

Estadístico	gl	p
4.29	1	0.0383

Cocientes de chance (odds ratio) de Mantel-Haenszel

Estadístico	Estim	LI 95%	LS 95%
MH Odds Ratio(1/2)	0.49	0.32	0.75
MH Odds Ratio(2/1)	2.04	1.33	3.12

En este ejemplo, no se detecta asociación significativa entre la aprobación del examen de ingreso y si el alumno se preparó solo o en academia a partir de la tabla marginal (sin controlar por facultad), el valor *p* del estadístico Chi cuadrado de Pearson es 0.3286, mayor al nivel de significación α=0.05, por tanto no hay evidencias para rechazar la hipótesis nula de independencia entre la aprobación del examen de ingreso y el modo de preparación para el mismo llevado a cabo por el alumno. El intervalo de confianza del cociente de chance para la tabla marginal incluye al valor 1, por tanto se debe concluir que la chance de aprobar el examen si el alumno se preparó en una academia es la misma que si se preparó solo.

Las medidas de asociación obtenidas sobre tablas marginales pueden conducir a interpretaciones falsas si la variable Z tiene efecto significativo sobre X, Y o ambas. En este ejemplo, controlando por facultad (Z), la prueba de hipótesis de Cochran-Mantel-Haenzel sugiere la existencia de asociación significativa (*p*=0.0383) y el cociente de chance Mantel-Haenzel (MH odds ratio 1/2) cuyo valor es 0.49, sugiere que la chance de aprobar el examen

si se preparó en academia es aproximadamente el doble de la chance de aprobar el examen si se preparó solo.

Las asociaciones obtenidas a través de tablas parciales (no mostradas en la tabla anterior) son conocidas con el nombre de asociaciones condicionales, ya que estudian la asociación X e Y condicional a un valor fijo de Z. En este ejemplo, para la facultad 1 la asociación existe pero para la facultad 2 no es significativa.

Regresión logística

Menú ESTADÍSTICAS ⇒ DATOS CATEGORIZADOS ⇒ REGRESIÓN LOGÍSTICA, permite modelar la relación entre una variable respuesta de naturaleza *dicotómica* en relación a una o más variables independientes o regresoras. Los coeficientes de la combinación lineal que modela esta relación permiten estimar la razón de productos cruzados (*odd ratio*) para cada variable regresora.

El modelo de regresión logística puede ser usado para predecir la probabilidad (p_i) de que la variable respuesta asuma un valor determinado, por ejemplo, probabilidad de éxito (y=1) en una variable dicotómica que asume los valores 0 y 1.

Para una respuesta binaria, el modelo de regresión logística simple, es decir con una regresora, tiene la siguiente forma:

$$Logit(p_i) = \log(p_i/(1-p_i)) = \alpha + \beta X_i$$

donde p_i es la probabilidad de éxito dado X_i, α es la ordenada al origen (constante), β es la pendiente o coeficiente de regresión asociado a X y X es la variable explicatoria. Luego, en regresión logística, se modela la transformación *Logit* de la probabilidad de éxito como una función lineal de una o más variables explicatorias.

El modelo logístico puede ser visto en el marco de una clase más general de modelos, donde se establece un modelo lineal para g(μ), siendo g(μ) una función del valor esperado de la variable respuesta y *g* una función conocida como función de enlace. En regresión logística el enlace canónico corresponde a la función logit del valor esperado *p* de la variable aleatoria observada la cual tiene distribución binomial (Hosmer y Lemeshow, 1989; Seber y Wild, 1989).

Si se simboliza con η al predictor lineal, en el caso de la regresión simple, $\eta = \alpha + \beta X_i$. La probabilidad de éxito será estimada por:

$$\hat{p} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}}$$

InfoStat permite ajustar modelos de regresión logística donde existe una variable respuesta binaria y una o más regresoras las cuales pueden ser continuas o categóricas. Si una variable categórica tiene más de dos categorías se deberá usar el generador de variables auxiliares

(*dummy)*, (ver Manejo de datos), ya que InfoStat asume que las regresoras categóricas son binarias.

InfoStat calcula para cada una de las variables del modelo el coeficiente de regresión, su error estándar, la estimación de la razón de productos cruzados (odd ratios), su intervalo de confianza, $-2(L_0-L_1)$ y el valor *p* para la prueba de hipótesis H_0: $\beta_i=0$ versus $\beta_i\neq0$.

También se presenta el logaritmo de la verosimilitud para el modelo elegido (L). La columna $-2(L_0-L_1)$ contiene –2 veces la diferencia de los logaritmos de la verosimilitud entre el modelo reducido (L_0) y el modelo completo (L_1). El modelo reducido, para la fila i-ésima de la tabla, es el modelo especificado por el usuario (L) sin la regresora correspondiente a la fila *i*. Por lo tanto, esta columna constituye la prueba del cociente de máxima verosimilitud para la hipótesis de coeficiente de regresión nulo para la variable en dicha fila. La prueba de hipótesis H_0: $\beta_i=0$ se realiza a partir del estadístico Chi cuadrado con un grado de libertad. InfoStat permite también guardar los residuos y los valores predichos para la evaluación del modelo ajustado.

Ejemplo 30: Se presenta a continuación un ejemplo en el que se estudia el efecto de la edad, la pérdida de peso inicial, el % del peso normal (PPI), el sexo (1:Masculinos, 0:Femeninos) y una medida del estado general del paciente (PSPI) sobre la sobrevida en pacientes con cáncer de pulmón evaluada a los tres meses de iniciado el tratamiento (datos: gentileza Dra. Norma Pilnik, Hospital Tránsito Cáceres de Allende, Córdoba). Los datos se encuentran en el archivo Logística.

Se realiza el ajuste de un modelo de regresión logística múltiple de la variable "muerto" (vale 0 si el paciente está vivo y 1 si el paciente falleció) en relación a la edad, sexo, PPI y PSPI. Las variables involucradas en el análisis se declaran en la ventana **Análisis de regresión logística**, en **Variable dependiente** se incluye la variable "muerto" y como **Variables Regresoras** "edad", "sexo", "PPI" y "PSPI".

Tabla 50: Análisis de regresión logística. Archivo Logística.

Análisis de Regresión Logística

Predictor	Coeficiente	EE	Odd	LI	LS	-2(L0-L1)	p
constante	-7.59	2.71	5.1E-04	2.5E-06	0.10	11.80	0.0006
EDAD	0.03	0.04	1.03	0.96	1.11	0.74	0.3905
PPI	0.12	0.05	1.12	1.02	1.24	5.46	0.0194
SEXO	-0.37	0.86	0.69	0.13	3.75	0.17	0.6788
PSPI	0.88	0.52	2.41	0.87	6.67	3.11	0.0776
Log-Likelihood						-36.57	___

De acuerdo a los resultados anteriores, la pérdida de peso inicial (PPI) es la única variable que muestra una relación significativa con la sobrevida del paciente. PSPI se insinúa como una posible predictora significativa para un nivel de significación $\alpha=0.10$.

Si la variable de respuesta es continua y se la desea transformar en binaria, en la ventana de diálogo que aparece después del selector de variables, se puede indicar un valor umbral para transformar la variable respuesta en binaria. En la parte inferior de la ventana donde dice **Considerar éxitos a valores** se habilita un campo para ingresar el umbral (valor numérico)

a partir del cual la respuesta se considerará éxito; para éxitos InfoStat hará *y*=1 y para fracasos *y*=0. En la misma ventana se puede indicar si desea guardar los residuos y/o los valores predichos por el modelo.

Sobrevida de Kaplan-Meier

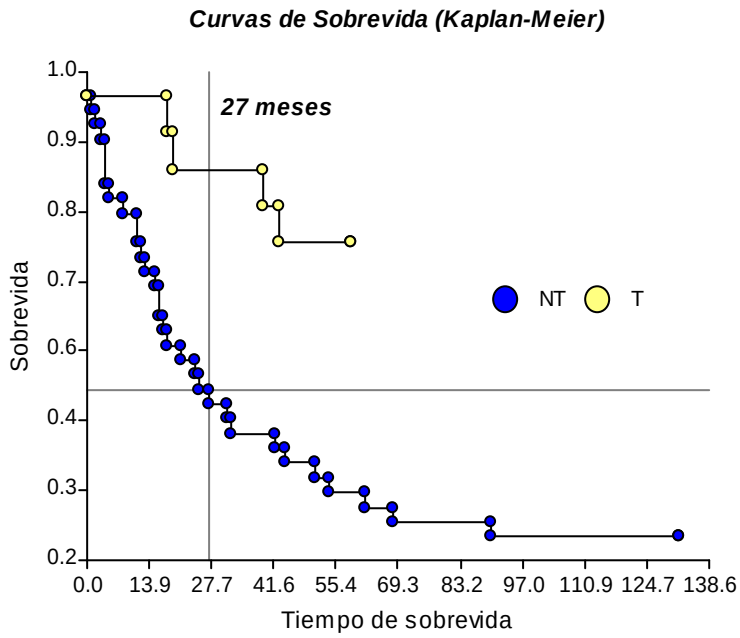
Menú ESTADISTICAS ⇒ DATOS CATEGORIZADOS ⇒ SOBREVIDA DE KAPLAN-MEIER , realiza un gráfico de curvas de sobrevida según el algoritmo de Kaplan y Meier y calcula los errores estándares que aparecen en la tabla de sobrevida según la descripción dada por Altman (1991).

El análisis de curvas de Kaplan-Meier permite estudiar la sobrevida de entidades en función de una variable independiente dicotómica (viva o muerta). InfoStat calcula también el valor del estadístico denominado *Log Rank,* para la prueba de igualdad de *k* >=2 curvas de sobrevida. Así un valor de *Log Rank* elevado presenta una correspondencia con un valor *p* pequeño, menor que un cierto nivel de significación previamente especificado, por ejemplo del 5%, lo cual indica que la al menos una de las k curvas de sobrevida comparadas es distinta.

La planilla de datos tiene que tener por lo menos dos columnas: una indicando el tiempo de sobrevida y la otra para indicar la variable independiente que refiere al estado en que se encuentra el individuo, esto puede ser: 0: vivo, 1: muerto por razones específicas, 2: muerto por otras causas que no son de interés, abandono del protocolo o perdida del sujeto. También pueden procesarse archivos de datos donde la variable independiente es dicotómica: 0: vivo y 1: muerto. Si existen dos o más grupos (digamos *k*) de sujetos se indica en una tercer columna que actuara como factor de clasificación con *k* niveles.

Ejemplo 31: Se presenta a continuación un ejemplo en el que se estudió a 56 pacientes con mieloma múltiple para los cuales se analizó la sobrevida (en meses) bajo dos condiciones: pacientes con transplante de médula (T) y no transplantados (NT). Para ellos se registró también su condición de vivo o muerto (Código de Censura: 0: vivo y 1: muerto). Los datos son gentileza de la Dra. Verónica Ortiz Corbella, Hospital Allende, Córdoba y se encuentran en el archivo Sobrevida.idb.

Se invocó el Menú ESTADISTICAS ⇒ DATOS CATEGORIZADOS ⇒ SOBREVIDA DE KAPLAN-MEIER, en la ventana **Sobrevida de Kaplan-Meier** se designó a "Sobrevida (meses)" como **Tiempo de sobrevida**, "Código de Censura" como **Criterio de Censura** y "Transp" como **Criterio de Clasificación (opcional)**. Los resultados fueron:



El valor p=0.001345 es menor que el nivel de significación α=0.05 que investigador requiere para la prueba, esto indica que la diferencia de sobrevida entre los grupos (NT y T) es significativa.

El grupo de los no transplantados (NT) presenta una curva de sobrevida que decrece más rápidamente que la de los transplantados (T).

En este gráfico se puede observar que el 50% de los no transplantados vive hasta los 27 meses aproximadamente.

Figura 21: Curvas de sobrevida de Kaplan-Meier. Archivo Sobrevida.idb.

Curvas de sensibilidad-especificidad

Menú DATOS CATEGORIZADOS ⇒ CURVAS DE SENSIBILIDAD-ESPECIFICIDAD, permite obtener *Curvas ROC* empíricas, gráficos de *sensibilidad* y de *especificidad* (por separado o simultáneos) y gráficos de valores predictivos positivos y/o negativos. Usando la siguiente tabla se definirán los cálculos que realiza InfoStat:

	Condición (+)	***Condición (-)***
Variable pronóstico (+)	Positivos Verdaderos	Falsos Positivos
Variable pronóstico (-)	Falsos Negativos	Negativos Verdaderos

Nota: **Condición** se refiere al estado verdadero de la entidad que se clasifica que puede ser positiva o negativa. El usuario puede ingresar esta condición o definirla en función de algún valor (mediana, media u otro) de la variable seleccionada para determinar la condición.

Variable pronóstico, se refiere a una variable cuyos valores (mayores que… o menores que…) llevan a pronosticar la condición.

$$Sensibilidad = \frac{Positivos\,Verdaderos}{Positivos\,Verdaderos + Falsos\,Negativos} \times 100\,;$$

$$Especificidad = \frac{Negativos\ Verdaderos}{Negativos\,Verdaderos + Falsos\,Positivos} \times 100\,;$$

$$Valor\ predictivo\ positivo\ = \frac{Positivos\,Verdaderos}{Positivos\,Verdaderos + Falsos\,Positivos} \times 100\,;$$

$$Valor\ predictivo\,negativo = \frac{Negativos\,Verdaderos}{Negativos\,Verdaderos + Falsos\,Negativos} \times 100\,.$$

Las curvas *ROC* se construyen graficando la *sensibilidad* (eje Y) vs. 1-*especificidad* (eje X).

Usando las curvas de sensibilidad y especificidad, superpuestas en un mismo gráfico, se puede determinar un punto de corte comúnmente llamado "threshold" (valor en el que se cruzan ambas curvas) para determinar el valor de la variable predictora (pronóstica) para el cual la sensibilidad y la especificidad se igualan.

Ejemplo 32: Se presenta a continuación un ejemplo en el que se estudiaron pacientes con tumores, todos con transplante de médula. En estos pacientes se registraron diferentes parámetros sanguíneos entre ellos los linfocitos a los 15 días del transplante ("Linfo15"). Se registró también la sobrevida ("Sobrevida (meses)"). Usando la mediana de la sobrevida se obtuvo la condición del paciente (0: sobrevida menor a la mediana y 1: sobrevida mayor a la mediana). El objetivo fue determinar el valor "threshold" de linfocitos (valor en que la especificidad es igual a la sensibilidad). Los datos son gentileza de la Dra. Verónica Ortiz Corbella, Hospital Allende, Córdoba y se encuentran en el archivo Sensibilidad.idb.

Se invocó el Menú APLICACIONES ⇒ OTROS ⇒ CURVAS DE SENSIBILIDAD-ESPECIFICIDAD, en la ventana **Especificidad, sensibilidad y valores predictivos** se designó a "Sobrevida (meses)" como **Condición**, y "Linfo15" como **Variables pronósticas** (si se especifica más de una variable pronóstica los cálculos se realizan para cada variable independientemente). En la siguiente ventana, en el espacio reservado para especificar la condición se activó la opción **Valores mayores o iguales que la mediana.** En el espacio

reservado para especificar las categorías de pronóstico se indicó que se consideren positivos los pacientes con valores de linfocitos mayores o iguales a cada uno de los valores observados. Se solicitaron simultáneamente las curvas de sensibilidad y especificidad. Se obtuvo el gráfico se presenta a continuación; en el mismo se determinó que el punto de corte de ambas curvas o "threshold" es aproximadamente 550 cel/µL.

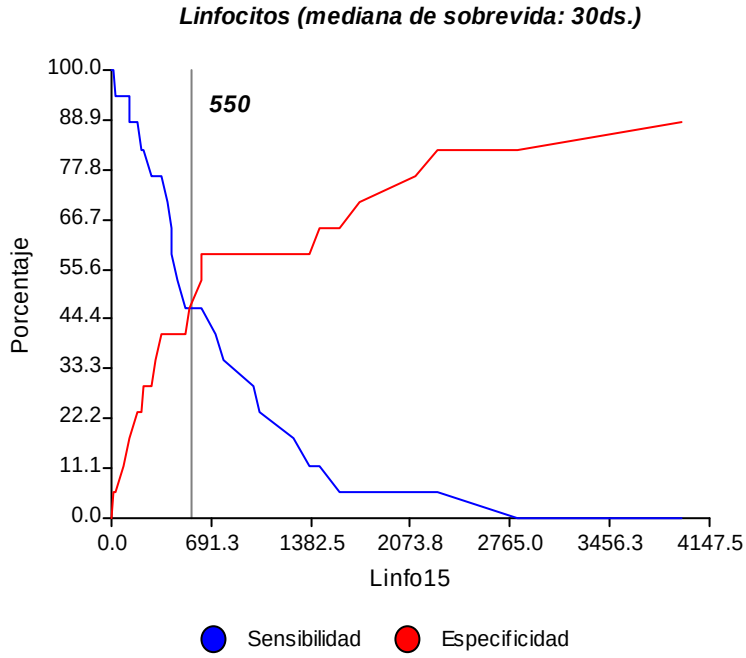


Figura 22: Curvas de sensibilidad-especificidad para determinación del punto de corte del parámetro linfocitos a los 15 días del transplante de médula en pacientes con tumores. Archivo Sensibilidad.idb.

Análisis multivariado

La estadística multivariada es usada para describir y analizar observaciones multidimensionales obtenidas al relevar información sobre varias variables para cada una de las unidades o casos en estudio. El módulo de Análisis Multivariado pone a disposición del usuario una serie de técnicas de análisis apropiadas para tablas de datos que contienen dos o más variables respuesta (columnas de la tabla) por cada caso (filas de la tabla). Luego, la organización de datos para un análisis multivariado se realiza en forma de una matriz con *n* filas (casos) conteniendo las *p* características (variables) registradas sobre un mismo individuo (datos de *p* variables observadas en cada uno de *n* casos, se colectan en una matriz $\mathbf{X}$, $n \times p$). La versión InfoStat/Estudiantil solo permite manejar un total de 9 columnas.

Tabla 51: Organización de datos multivariados.

Individuos	V_1	V_2		V_j		V_p
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1p}
2	X_{21}	X_{22}	...	X_{2j}	...	X_{2p}
.	.	.	...	.	...	.
.	.	.	...	.	...	.
n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{np}

Cada observación multivariada es representada por un vector *p*-dimensional de variables aleatorias y se puede conceptualizar como un punto en R^p, con coordenadas igual al valor asumido por cada una de las variables. Si se tienen 3 individuos y 2 variables aleatorias (por ejemplo altura y peso) registradas sobre cada uno de los individuos, asumiendo los valores que se muestran a continuación, la representación de las tres observaciones bivariadas puede, en el espacio de las dos variables) ser la siguiente:

Caso	Altura	Peso
1	2	3
2	1	2
3	3	2

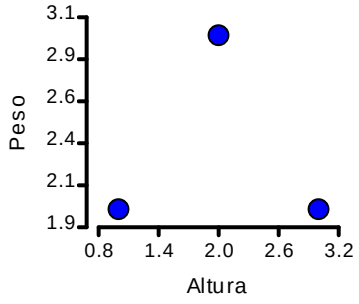
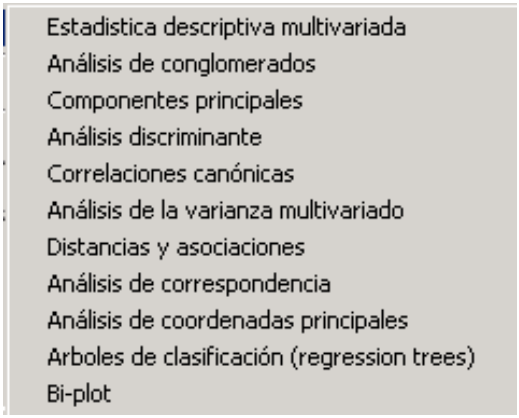


Figura 23: Representación de datos multivariados.

Cuando más de tres variables son relevadas para cada caso la visualización directa de las observaciones no es posible, por ello se utilizan técnicas de reducción de dimensión y proyecciones de la nube de puntos que representan las observaciones en espacios de fácil visualización como es el plano. Gráficos comúnmente utilizados para visualizar y comparar observaciones multivariadas son los *gráficos de estrellas*, las *matrices de diagramas de dispersión*, los gráficos de *perfiles multivariados* (ver Capítulo Gráficos).

InfoStat permite aplicar técnicas de análisis para entender la relación entre variables medidas simultáneamente, comparar, agrupar y/o clasificar observaciones en función de varias variables o variables en función de observaciones. Cuando se elige Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO, se visualiza el siguiente submenú.

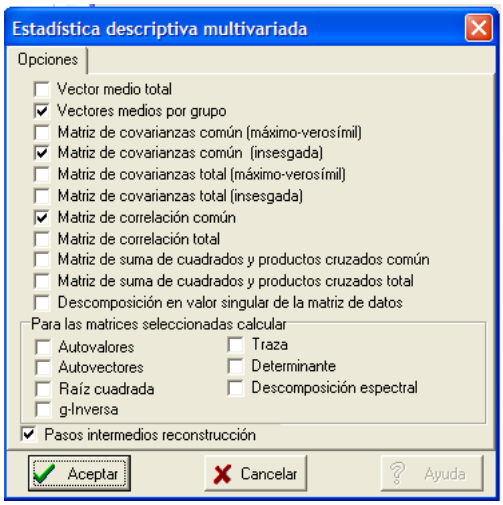


Al decidir el tipo de análisis multivariado a usar, aparece una ventana que permitirá indicar las variables a analizar y establecer, de ser necesario, un criterio de clasificación.

Estadística descriptiva multivariada

Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ ESTADÍSTICA DESCRIPTIVA MULTIVARIADA. En el selector de variables de la ventana **Estadística descriptiva multivariada** especificar las variables respuesta en **Variables**. El **Criterio de clasificación** es opcional. De existir una o más columnas de la tabla de datos que clasifiquen las observaciones en grupos, ellas pueden señalarse como criterios de clasificación para reducir la dimensión de la matriz de observaciones. En tal caso InfoStat usará como unidad de análisis cada uno de los grupos formados por el criterio de clasificación.

Para el cálculo de estadísticos descriptivos, el número de observaciones utilizado por InfoStat corresponde al número de casos activos. Si se desea eliminar del análisis los casos con al menos una variable con datos faltantes se debe activar la casilla **Eliminar Registros Incompletos** ubicada al pie de la ventana de **Estadística descriptiva multivariada** (opción por defecto). InfoStat provee automáticamente el/los vector/es medios por grupo, la matriz de covarianza insesgada y la matriz de correlación. Otros estadísticos que pueden ser seleccionados, activando el casillero correspondiente en la ventana de opciones, se presentan a continuación:



Nociones teóricas sobre estadística descriptiva multivariada

La descripción de muestras aleatorias multivariadas se puede realizar mediante el cálculo de estadísticos muestrales. Johnson y Wichern (1998) definen un muestreo aleatorio, de observaciones multivariadas, como aquel donde: 1) las mediciones tomadas sobre casos diferentes no se encuentran correlacionadas y 2) la distribución conjunta de las *p* variables es la misma para cada caso. A continuación se describen algunos estadísticos descriptivos multivariados frecuentemente usados para la descripción de muestras aleatorias multivariadas.

Vectores medios (total y por grupo): para cada variable de la matriz de datos se calcula la media muestral. Si la matriz de datos es de dimensión $n \times p$ habrá p medias muestrales, denotadas por $\bar{x}_j$ con j=1,...,p, cada una obtenida como:

$$\bar{x}_j = \sum_{i=1}^{n} \frac{x_{ij}}{n} \quad \text{para j=1,...,}p$$

donde x_{ij} es el i-ésimo valor de la j-ésima variable. Las p medias constituyen en este caso el **vector de medias total**. Si se indicó un criterio de clasificación se podrá obtener el **vector de medias por grupo** conformado a partir de las medias de las p variables calculadas a partir de las observaciones de cada grupo.

Matrices de varianza-covarianza: la varianza muestral, calculada a partir de las n mediciones sobre cada variable, será denotada por S_j^2. Para una matriz de datos de dimensión $n \times p$ habrá p varianzas muestrales, cada una obtenida a partir de la expresión:

$$S_j^2 = c\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \quad \text{para j=1,...,}p$$

donde la constante c puede ser $1/n$ o $1/(n-1)$ de acuerdo se trate del estimador máximo verosímil o del estimador insesgado de la varianza poblacional, respectivamente.

La covarianza muestral mide asociación lineal entre dos variables, es decir mide cómo varían dos variables conjuntamente. La covarianza entre la variable j-ésima y la variable k-ésima es obtenida por la siguiente expresión:

$$S_{jk} = c\sum_{i=l}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k) \quad \text{para } j,k=1,..., p$$

Al solicitar una matriz de varianzas y covarianzas InfoStat dispone las varianzas y las covarianzas de las *p* variables en una matriz cuadrada *p*×*p* simétrica. Esta matriz contiene las varianzas de cada una de las *p* variables sobre la diagonal principal y las covarianzas entre cada par de variables como elementos fuera de la diagonal principal. Luego, si se denota como **S** a dicha matriz, esta tendrá la siguiente forma:

$$S = \begin{bmatrix} S_1^2 & S_{12} & \ldots & S_{1p} \\ S_{21} & S_2^2 & \ldots & S_{2p} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ S_{p1} & S_{p2} & \ldots & S_p^2 \end{bmatrix}$$

La matriz de varianzas-covarianzas tiene *p* varianzas y [*p*(*p*-1)]/2 covarianzas. En la **matriz de varianza-covarianza máximo verosímil** la constante *c* usada en el cálculo de cada elemento será 1/*n*, y en la **matriz de varianza-covarianza insesgada**, la constante *c* será 1/(*n*-1). En la **Matriz de covarianzas total** (en cualquiera de sus dos versiones), todas las *n* observaciones son usadas en el cálculo de varianzas y covarianzas.

En la **Matriz de covarianzas común** (en cualquiera de sus dos versiones), la matriz resultante se obtiene a partir del *promedio ponderado* de las matrices de varianzas y covarianzas para cada grupo. Luego, si existe una variable clasificatoria que separa la matriz de datos en dos o más grupos, la matriz de covarianza común será obtenida a través del promedio ponderado de las matrices de covarianzas estimadas en cada grupo separadamente. Por ejemplo si se tienen dos grupos, la matriz de covarianzas común será:

$$\mathbf{S}_{comun} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

donde $\mathbf{S}_1$ y $\mathbf{S}_2$ son las matrices de covarianza del grupo 1 y del grupo 2 estimadas con grados de libertad $(n_1 - 1)$ y $(n_2 - 1)$, respectivamente. Esta matriz de covarianza común tiene sentido cuando no existen diferencias entre las matrices de covarianzas de cada grupo.

Matriz de correlación total: es una matriz cuadrada *p*×*p* simétrica conteniendo el valor 1 sobre la diagonal principal y los coeficientes de correlación de Pearson entre cada par de variables como elementos fuera de la diagonal principal. El *coeficiente de correlación producto momento de Pearson* es una medida de la magnitud de la asociación lineal entre dos variables que no depende de las unidades de medida de las variables originales. Para las variables j-ésima y k-ésima se define como:

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_j^2 \; S_k^2}} = \frac{\left(\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k) \right)}{\sqrt{\left(\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2 \right) \quad \left(\sum_{i=1}^{n} (x_{ik} - \overline{x}_k)^2 \right)}}$$

El coeficiente tiene el mismo valor cuando S_{jk}, S_j^2 y S_k^2 son expresadas con divisor *n* o (*n*-1). El coeficiente de correlación muestral representa la covarianza de los valores muestrales estandarizados. Asume valores en el intervalo [-1;1] y el signo indica la dirección de la asociación (valores negativos se producen cuando la tendencia promedio indica que cuando un valor en el par observado es más grande que su media, el otro valor es más pequeño que su media).

Matriz de correlación común: es una matriz cuadrada *pxp* simétrica conteniendo el valor uno sobre la diagonal principal y los coeficientes de correlación de Pearson entre cada par de variables como elementos fuera de la diagonal principal. A diferencia de la matriz de correlación total, los coeficientes son calculados después de corregir por el efecto de grupo (grupo definido según un criterio de clasificación). Esta matriz corresponde a la matriz de correlaciones parciales de cada par de variables, ajustadas por el efecto de grupo. Por ejemplo, si una matriz de datos de altura y peso de individuos contiene observaciones de dos grupos y no se tiene en cuenta esta clasificación, se podría obtener una correlación negativa cuando en realidad esta es positiva en ambos grupos (ver Figura 24). Luego, es necesario considerar que hay dos grupos con medias diferentes para obtener correlaciones positivas entre las dos variables. Este tipo de correlaciones es conocido como *correlaciones parciales* y son las que InfoStat reporta al pedir la matriz de correlación común.

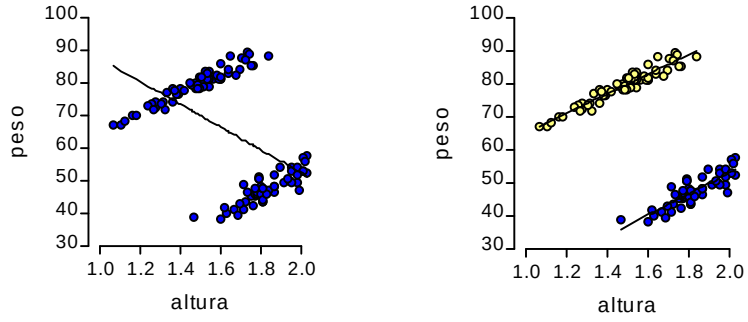


Figura 24: Gráficos de dispersión de las variables altura y peso para dos grupos de individuos.

Matriz de sumas de cuadrados y productos cruzados: es una matriz simétrica *pxp* construida de la misma forma que **S** pero haciendo *c*=1. Las *sumas de cuadrados* en la diagonal principal y las *sumas de productos cruzados* fuera de ella, son por sí mismas importantes estadísticos muestrales. Esta matriz puede construirse a partir de las *n* observaciones (matriz total) o como un promedio ponderado de las matrices correspondientes por grupo (matriz común), como fue explicado anteriormente. Algebraicamente:

$$\boldsymbol{W} = (n\text{-}1)\,\boldsymbol{S} = \begin{bmatrix} x_{11}-\overline{x}_1 & x_{21}-\overline{x}_1 & \dots & x_{n1}-\overline{x}_1 \\ x_{12}-\overline{x}_2 & x_{22}-\overline{x}_2 & \dots & x_{n2}-\overline{x}_2 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{1p}-\overline{x}_p & x_{2p}-\overline{x}_p & \dots & x_{np}-\overline{x}_p \end{bmatrix} \times \begin{bmatrix} x_{11}-\overline{x}_1 & x_{21}-\overline{x}_2 & \dots & x_{1p}-\overline{x}_p \\ x_{21}-\overline{x}_1 & x_{22}-\overline{x}_2 & \dots & x_{2p}-\overline{x}_p \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1}-\overline{x}_1 & x_{2n}-\overline{x}_2 & \dots & x_{np}-\overline{x}_p \end{bmatrix}$$

Descomposición en valor singular de la matriz de datos: es una forma de describir una matriz de datos. InfoStat permite aplicar esta descomposición a la tabla de datos activa o a matrices de covarianza, correlación y/o sumas de cuadrados. La función aplicada a una matriz permite obtener el conjunto de *autovalores* y *autovectores* que la describen. Una matriz rectangular $\mathbf{X}$ de dimensiones $n \times p$ puede ser escrita en términos de su descomposición por valor singular de la siguiente manera, $\mathbf{X} = \mathbf{UDV}'$, donde $\mathbf{U}$ es $n \times p$ con columnas ortogonales, $\mathbf{V}$ es una matriz ortogonal $p \times p$ y $\mathbf{D}$ una matriz diagonal $p \times p$, cuyos elementos son llamados valores singulares de $\mathbf{X}$.

La *descomposición espectral* de una matriz puede verse como un caso particular de descomposición, que se aplica a matrices cuadradas y simétricas, en tal caso $\mathbf{U}$ y $\mathbf{V}$ son iguales. InfoStat permite aplicar este tipo de descomposición a cualquiera de las matrices que se listan en el menú Estadística descriptiva multivariada. Con la opción **Pasos intermedios de la reconstrucción** provee , según el orden especificado por el usuario, *r*, la reconstrucción de la matriz que se está descomponiendo a partir de *r* pares de autovalores y autovectores. El usuario puede visualizar la reconstrucción paso a paso. Cuando *r* es igual al número de autovectores distintos de cero, el último paso de reconstrucción devuelve la matriz original.

Autovectores y autovalores de $\mathbf{S}$: ambos resumen la información en términos de variabilidad. Los *autovectores* constituyen un conjunto de vectores bases para graficar los datos y los *autovalores* representan la variabilidad de los datos en cada una de las direcciones dadas por los autovectores. Luego, los autovalores son medidas de variabilidad, mientras que los autovectores expresan la dirección de la variabilidad.

Determinante: el determinante de $\mathbf{S}$ también conocido como *varianza generalizada* es una forma de resumir variabilidad multivariada ya que resume información sobre todas las varianzas y covarianzas entre las variables en un solo número. Calculado sobre $\mathbf{S}$ también puede ser visto como el producto de sus autovalores.

Muchas matrices de covarianza pueden tener el mismo valor de varianza generalizada. La varianza generalizada es cero cuando al menos un vector desviación $\left(\mathbf{x}_j - \overline{\mathbf{x}}_j\right)$ es una combinación lineal de los otros. Un resultado importante establece que si $p \geq n$ *(*más variables que observaciones) el determinante de la matriz de covarianza será cero. El determinante de la matriz de correlación $\mathbf{R}$ es la varianza generalizada de las variables estandarizadas.

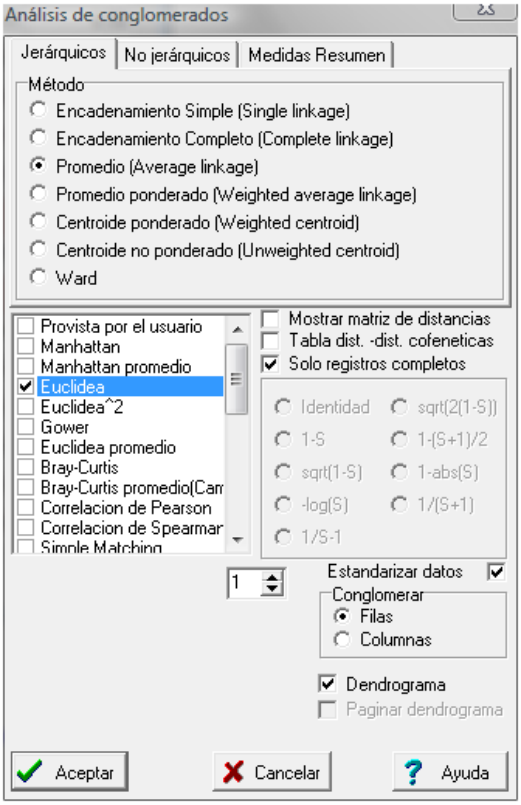
Traza: es la suma de los elementos diagonales de una matriz. Calculado sobre $\mathbf{S}$ es otra medida unidimensional de variabilidad multivariada. La varianza muestral total es la suma

de las varianzas de cada variable, *i.e.* traza(**S**). No tiene en cuenta la estructura de correlación. Sobre **S** también puede ser vista como la suma de sus autovalores.

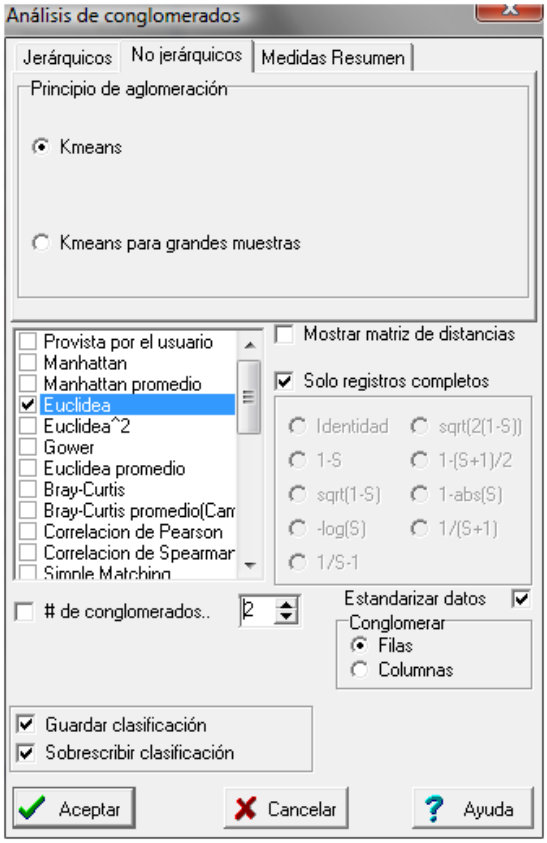
Raiz cuadrada: Devuelve la matriz R tal que R*R es igual a la matriz seleccionada. La obtención de R se basa en el algoritmo de la descomposición por valor singular.

G-inversa: Devuelve una inversa generalizada de la o las matrices seleccionadas previamente. Los elementos de la diagonal de la *inversa de la matriz de correlación* dan idea de la magnitud en la que las variables son funciones lineales de otras. Tales elementos diagonales suelen llamarse factores de *inflación de varianza*. El j-ésimo elemento diagonal es $1/(1 - R_j^2)$ donde R_j^2 es el cuadrado del coeficiente de determinación de la regresión múltiple entre la j-ésima variable y las otras variables. Un elemento de la diagonal grande indica que la variable correspondiente está altamente correlacionadas con alguna de las otras variables. Cuando una o más variables es una función lineal de otras de las variables consideradas, la matriz de correlación no es de rango completo.

Análisis de conglomerados



Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DE CONGLOMERADOS, permite implementar distintos procesos para agrupar objetos descriptos por un conjunto de valores de varias variables. Los objetos generalmente representan las filas de la tabla de datos. Ocasionalmente, estos procedimientos son usados para agrupar variables en lugar de observaciones (es decir conglomerar columnas en lugar de filas). En InfoStat, la ventana **Análisis de conglomerados** permite seleccionar las variables del archivo que se usarán en el análisis e indicar una o más variables como criterio de clasificación con el objetivo de resumir varios registros en un único caso. Al presionar el botón **Aceptar** aparece otra ventana llamada **Análisis de conglomerados** la cual tiene tres solapas: **Jerárquicos, No jerárquicos** y **Medidas resumen.** En caso que se haya indicado un criterio de clasificación de registros, en la solapa **Medidas de resumen**, InfoStat permite escoger entre medidas de posición como la **media**, **mediana**, **mínimo**, **máximo** y de dispersión como la **varianza** y **desviación estándar** para resumir la información de cada variable en cada conjunto de registros definido por el criterio de clasificación (por defecto usa la media).



En la solapa **Jerárquicos** y **No jerárquicos**, se puede elegir el **método** (por defecto se selecciona automáticamente el agrupamiento promedio entre los jerárquicos o *K-Means* como algoritmo no jerárquico), y el tipo de **distancia** (por defecto Euclídea promedio) a utilizar en la conformación de conglomerados. InfoStat permite activar la celda **estandarizar los datos**, esta opción estandariza automáticamente cada columna seleccionada como variable antes de realizar el agrupamiento. El análisis puede realizarse por **filas**, en tal caso se agruparán registros o por **columnas** para formar conglomerados de variables. Cuando el número de objetos a clasificar es grande, se puede dividir al dendrograma en páginas (opción paginar dendrograma)

Tanto para los conglomerados no jerárquicos como para los jerárquicos, cuando se está agrupando casos (conglomerar filas) o variable (conglomerar columnas), mediante la activación del casillero **Guardar clasificación**, InfoStat genera una nueva columna en la tabla de datos activa que contiene la designación del número de grupo al que fue asignada cada observación. El número de grupos debe ser especificado de antemano en el casillero **Número de conglomerados**. InfoStat provee automáticamente de un gráfico indicando la reducción en la función objetivo del agrupamiento, en relación al número de conglomerados (desde dos hasta el número indicado por el usuario), identificando los grupos formados con diferentes colores. En el caso de conglomerados no jerárquicos, el número recomendado de grupos es aquel que se asocia con una caída mayor de la función respecto al número inmediato anterior.

Para los conglomerados jerárquicos, InfoStat produce automáticamente el dendrograma correspondiente a la evolución del agrupamiento en función de la distancia seleccionada. La información visualizada en el dendrograma puede ser leída en la ventana **Resultados**.

Ejemplo 33: se realizó un plan de recolección de datos para analizar semejanzas y diferencias morfométricas entre 14 genotipos (cultivares) de garbanzo. Se midieron 9 variables, como son el largo, el ancho y el espesor de la vaina entre otras, sobre cada observación correspondiente a un genotipo. Existen varias observaciones por objeto que se desea agrupar. Los datos (gentileza Ing. Julia Carreras, Facultad de Ciencias Agropecuarias-U.N.C.), se encuentran en el archivo Garbanzo.

Para realizar un análisis de conglomerados jerárquico se eligió: ESTADÍSTICAS ⇒ ANALISIS MULTIVARIADO ⇒ ANÁLISIS DE CONGLOMERADOS. En la ventana **Análisis de conglomerados** se especificó como **Variables** a todas las mediciones y como **Criterios de clasificación** al "genotipo". En la solapa **Jerárquicos** se eligió el método de agrupamiento **Encadenamiento promedio (average linkage)** y la distancia **Euclídea**. Se seleccionó el campo **Estandarizar datos** y se solicitó un análisis de aglomeración de **Filas**. En número de conglomerados se puso 4 y para resumir las observaciones de un mismo genotipo se utilizó la opción seleccionada por defecto en la solapa **Medidas resumen** (**Media**). Se obtuvo el siguiente dendrograma:

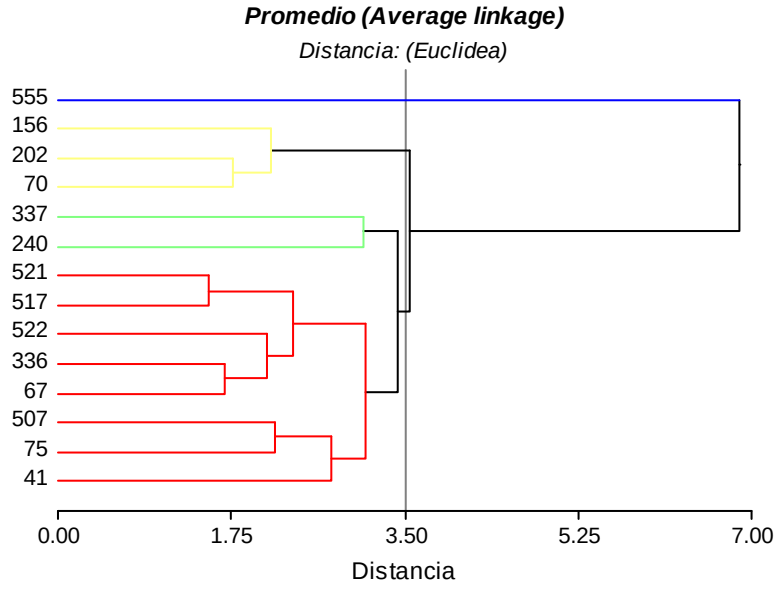


Figura 25: Dendrograma. Archivo Garbanzo.

En este ejemplo, fijando un criterio de corte arbitrario en la distancia 3.5, el genotipo 555 se separa del resto. Los genotipos 156, 202 y 70 forman un grupo, los genotipos 337 y 240 otro grupo y los restantes genotipos conforman otro grupo. Es un criterio frecuentemente utilizado trazar la línea de referencia a una distancia igual al 50% de la distancia máxima (en este caso la distancia máxima es cercana a 7, por lo que el punto de corte se trazó en 3.5).

Nociones teóricas sobre el análisis de conglomerados

El agrupamiento de objetos multivariados es frecuentemente utilizado como método exploratorio de datos con la finalidad de obtener mayor conocimiento sobre la estructura de las observaciones y/o variables en estudio. Si bien es cierto que el proceso de agrupamiento conlleva inicialmente a una pérdida de información ya que se sitúan en una misma clase unidades que no son idénticas (solo semejantes), la síntesis de la información disponible sobre las unidades consideradas puede facilitar considerablemente la visualización de relaciones multivariadas de naturaleza compleja. Se recurre a técnicas de agrupamiento

cuando no se conoce una estructura de agrupamiento de los datos "a priori" y el objetivo operacional es identificar el agrupamiento natural de las observaciones. Las técnicas de clasificación basadas en agrupamientos implican la distribución de las unidades de estudio en clases o categorías de manera tal que cada clase (conglomerado) reúne unidades cuya similitud es máxima bajo algún criterio. Es decir los objetos en un mismo grupo comparten el mayor número permisible de características y los objetos en diferentes grupos tienden a ser distintos.

Para agrupar objetos (casos o variables) es necesario seguir algún algoritmo. La palabra algoritmo designa un conjunto de reglas operativas sistemáticas que permiten la realización de un tipo de tareas paso a paso para obtener un resultado. Los algoritmos o métodos de agrupamiento permiten identificar clases existentes en relación a un conjunto dado de atributos o características. En distintas áreas del conocimiento se encuentran estos algoritmos bajo diferentes nombres como son *clasificación automática*, análisis tipológico (del francés "analyse typologique"), *análisis de agrupamiento* (del inglés "cluster analysis"), *taxonomía numérica*, etc. Los algoritmos de clasificación pueden dividirse en *no jerárquicos* y *jerárquicos.* En las técnicas de clasificación no jerárquicas se desea obtener una única descomposición o partición del conjunto original de objetos en base a la optimización de una función objetivo. Mientras que en las técnicas de clasificación jerárquicas, se pretenden encontrar particiones jerarquizadas, esto es, consecutivamente más finas (o menos finas), luego los objetos son unidos (o separados) en grupos paso por paso. En biología (taxonomía) las técnicas jerárquicas son tradicionales ya que traducen mejor la complejidad de la organización de los seres vivos y la existencia de distintos niveles evolutivos. En otras aplicaciones las técnicas no jerárquicas pueden ofrecer una descripción apropiada de los datos, por ejemplo clasificación de libros en una biblioteca.

Los algoritmos de agrupamiento pueden ser *supervisados* o *no supervisados* según si el número de clases a ser obtenidas es fijado "a priori" por la persona que conduce el experimento o si éste resulta de la aplicación de la técnica de clasificación. Muchas veces, informaciones preliminares disponibles o resultados de experimentos pilotos, pueden orientar al experimentador o usuario en la selección del número de clases. Otras veces, se conoce algún valor máximo para el número de clases, y entonces el algoritmo se implementa especificando dicho valor y luego, en relación con los resultados obtenidos, se vuelven a realizar agrupamientos. Las técnicas de clasificación jerárquicas son generalmente del tipo *no supervisadas*.

El agrupamiento logrado dependerá no sólo del algoritmo de agrupamiento elegido sino también de la medida de distancia seleccionada, del número de grupos que deben ser formado (cuando esta información existe), de la selección de las variables para el análisis y del escalamiento de las mismas. Textos tradicionales que abordan la problemática asociada con la formación de conglomerados son los de Anderberg (1973) y de Everitt (1974).

En el análisis de conglomerados de casos o registros individuales se parte de una matriz de datos $n \times p$ (supongamos p mediciones o variables en cada uno de los n objetos estudiados), que luego es transformada en una matriz de distancia ($n \times n$) donde el elemento i,j-ésimo mide la distancia entre pares de objetos i y j para $i,j=1,...,n$. Los elementos de la matriz de distancia son funciones de distancias métricas o no métricas. En el análisis de

conglomerados de variables se usará una matriz de distancia ($p \times p$) donde el elemento i,j-ésimo mide la distancia entre pares de variables i y j para $i,j=1,...,p$.

Cuando se disponen de numerosas variables para realizar el agrupamiento, es común utilizar (antes del análisis de conglomerados) técnicas de reducción de dimensión tal como Análisis de Componentes Principales para obtener un número menor de variables capaces de expresar la variabilidad en los datos. Esta técnica puede facilitar la interpretación de los agrupamientos obtenidos.

En la práctica, se recomienda aplicar varios algoritmos de agrupamiento y de selección o combinación de variables para cada conjunto de datos. Seleccionando, finalmente, desde los agrupamientos realizados la interpretación más apropiada. InfoStat provee automáticamente el valor del *coeficiente de correlación cofenética* el cual puede ser usado para seleccionar uno de varios agrupamientos alternativos. Este coeficiente indica la correlación de las distancias definidas por la métrica de árbol binario con las distancias originales entre objetos, luego se espera que el agrupamiento con mayor coeficiente sea el que mejor describe el agrupamiento natural de los datos.

Es importante remarcar que los procedimientos de agrupamiento producen resultados exitosos cuando la matriz de datos tiene una estructura que es posible interpretar desde el problema que originó la recolección de la información. Por ello, logrados los grupos es importante caracterizar los mismos a través de diversas medidas resumen para favorecer la interpretación del agrupamiento final.

Métodos de agrupamiento jerárquicos

Los métodos jerárquicos producen agrupamientos de tal manera que un conglomerado puede estar contenido completamente dentro de otro, pero no está permitido otro tipo de superposición entre ellos. Los algoritmos de conglomeración jerárquicos utilizados con fines de agrupamiento pueden ser *aglomerativos* o *divisivos* (utilizan fusiones o divisiones sucesivas de los objetos a agrupar).

Los métodos aglomerativos realizan grupos por el procedimiento de uniones sucesivas. En el inicio hay tantos grupos como objetos. Los objetos similares se agrupan primero y esos grupos iniciales son luego unidos de acuerdo a sus similitudes. Los métodos jerárquicos divisivos comienzan asumiendo que todos los objetos pertenecen a un mismo grupo al cual particionan en subdivisiones cada vez más finas, hasta el punto donde cada objeto es considerado un conglomerado de tamaño unitario. InfoStat trabaja con métodos aglomerativos ya que estos son más satisfactorios con respecto a los tiempos de cálculo.

Los resultados de agrupamientos jerárquicos se muestran en un *dendrograma* (diagramas de árboles en dos dimensiones), en el que se pueden observar las uniones y/o divisiones que se van realizando en cada nivel del proceso de construcción de conglomerados (ver Figura 25). Las ramas en el árbol representan los conglomerados. Las ramas se unen en un nodo cuya posición a lo largo del eje de distancias indica el nivel en el cual la fusión ocurre. El nodo donde todas las entidades forman un único conglomerado, se denomina *nodo raíz.* Debido a que en cada nivel se evalúa la unión de dos observaciones (o dos conglomerados), estos dendrogramas se conocen como *árboles binarios*. En la práctica, el interés principal suele

estar centrado en resultados intermedios donde los objetos se encuentran clasificados en un número moderado de conglomerados.

Una de las principales características de los procedimientos de agrupamiento jerárquicos aglomerativos es que la ubicación de un objeto en un grupo no cambia, o sea, que una vez que un objeto se ubicó en un conglomerado, no se lo reubica. Este objeto puede ser fusionado con otros pertenecientes a algún otro conglomerado, para formar un tercero que incluye a ambos.

Los algoritmos aglomerativos proceden de la siguiente manera: inicialmente, cada objeto pertenece a un conglomerado diferente; en la siguiente etapa se fusionan los dos objetos más cercanos para formar el primer conglomerado; en la tercera etapa, un nuevo objeto se agrega al conglomerado formado en la primera etapa u otros dos objetos se fusionan formando un segundo conglomerado. El proceso continúa de manera similar hasta que, eventualmente, se forma un solo conglomerado que contiene todos los objetos como integrantes del mismo. Las técnicas de agrupamiento jerárquico difieren por la regla de asignación de objetos a un conglomerado o fusión de conglomerados que utilizan. A continuación se listan los algoritmos de agrupamiento jerárquicos disponibles en InfoStat:

Encadenamiento simple *(single linkage)* (Florek *et al.,* 1951a, 1951b): Los grupos se unen basándose en la distancia entre los dos miembros más cercanos. Este método también conocido con el nombre de procedimiento del vecino más cercano *(nearest neighbor)* utiliza el concepto de mínima distancia y comienza buscando los dos objetos que la minimizan. Ellos constituyen el primer conglomerado. En las etapas siguientes se procede como se ha explicitado en el punto anterior, pero partiendo de *n-1* objetos donde uno de ellos es el conglomerado formado anteriormente. La distancia entre conglomerados está definida como la distancia entre sus miembros más cercanos.

Ejemplo: se aplica la técnica de encadenamiento simple a partir de la siguiente matriz de distancias entre cinco individuos:

$$D_1 = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ 5 & 3 & 0 & & \\ 6 & 8 & 4 & 0 & \\ 8 & 7 & 11 & 2 & 0 \end{bmatrix} \end{array}$$

donde la i-ésima fila y la j-ésima columna dan la distancia d_{ij} entre los individuos *i* y *j* .

En la primera etapa se juntan A y B, ya que d(AB) = 1.0 es el menor elemento de D_1. Una vez formado este conglomerado, se calculan las distancias entre el cluster y el resto de los individuos C, D y E. Se obtienen de D_1 como sigue:

d(AB, C)= mín {d(AC), d(BC)} = d(BC) = 3.0

d(AB, D)= mín {d(AD), d(BD)} = d(AD) = 6.0

d(AB, E)= mín {d(AE), d(BE)} = d(BE)= 7.0

Se forma entonces la matriz de distancias D_2, cuyos elementos son distancias entre individuos y distancias entre individuos y grupo

$$D_2 = \begin{array}{c} \\ (AB) \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} (AB) & C & D & E \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 6 & 4 & 0 & \\ 7 & 11 & 2 & 0 \end{bmatrix} \end{array}$$

La menor entrada de D_2 es d(DE)= 2.0. Por lo tanto la decisión es agrupar los individuos D y E, que forman un nuevo grupo, no agregando ningún otro al conglomerado ya formado.

En la tercera etapa, se calculan las siguientes distancias:

d(AB, C) = 3.0

d(AB, DE) = min {d(AD), d(AE), d(BD), d(BE)} = d(AD) = 6.0

d(DE, C) = min {d(CD), d(CE)} = d(CD) = 4.0

Así, D_3 resulta:

$$D_3 = \begin{array}{c} \\ (AB) \\ C \\ (DE) \end{array} \begin{array}{c} \begin{array}{ccc} (AB) & C & (DE) \end{array} \\ \begin{bmatrix} 0 & & \\ 3 & 0 & \\ 6 & 4 & 0 \end{bmatrix} \end{array}$$

El menor elemento en D_3 es d(AB, C). Esto indica que el individuo C debería agruparse en el primer conglomerado (con A y B). En el último paso, los dos grupos se fusionan formando un único conglomerado (single), que contiene a todos los individuos.

Dado que el procedimiento de encadenamiento simple une conglomerados en función de la mínima distancia entre sus elementos, el procedimiento puede tener problemas cuando hay grupos muy cercanos o con cierta superposición. El procedimiento de encadenamiento simple, es uno de los pocos procedimientos de clasificación que tienen un buen desempeño con configuraciones de conglomerados no elípticas (datos en cadena). Este método es recomendado para detectar estructuras de agrupamiento irregular y elongadas. Tiende a separar los extremos de la distribución antes de separar los grupos principales (Hartigan, 1981). Por ello tiende a producir agrupamientos de grupos en cadena.

Encadenamiento completo *(complete linkage)* (Sorensen, 1948)*:* La distancia entre conglomerados es la del par de objetos más distantes. Este método también conocido con el nombre de vecino más lejano (*farthest neighbor*) es similar al anterior, pero las distancias se definen ahora, como la distancia entre pares de individuos más distantes. Para ejemplificar el procedimiento de encadenamiento completo se trabaja con la matriz de distancias D1 a partir de la cual se desarrolló la técnica de encadenamiento simple. Primero se fusionan los individuos A y B como en el encadenamiento simple, sin embargo las distancias entre este conglomerado y los tres individuos restantes se obtiene de D1 como sigue:

d(AB, C) = máx {d(AC), d(BC)} = d(AC) = 5.0

d(AB, D) = máx {d(AD), d(BD)}= d(BD) = 8.0

d(AB, E) = máx {d(AE), d(BE)}= d(AE) = 8.0

Esto da:

$$D_2 = \begin{matrix} AB \\ C \\ D \\ E \end{matrix}\begin{bmatrix} 0 & & & \\ 5 & 0 & & \\ 8 & 4 & 0 & \\ 8 & 11 & 2 & 0 \end{bmatrix}$$

Por lo tanto, los individuos D y E se unen en un nuevo conglomerado. Por cálculos similares se obtiene:

$$D_3 = \begin{matrix} AB \\ C \\ DE \end{matrix}\begin{bmatrix} 0 & & \\ 5 & 0 & \\ 8 & 11 & 0 \end{bmatrix}$$

Luego, C se une al conglomerado AB y finalmente ambos conglomerados se juntan.

Puede demostrarse que este método es idéntico al método conocido como "mínimo árbol" (Anderberg, 1973). El algoritmo para calcular la distancia entre conglomerados corresponde a la división del árbol formado mediante la unión, en primera instancia, de los dos objetos con menor separación, luego aquél con la próxima menor separación, etc., y que une finalmente aquellos con la mayor separación. Tiende a producir grupos con igual diámetro y es poco resistente a valores extremos (Milligan, 1980).

Encadenamiento promedio *(average linkage)* o **UPGMA (unweighted pair-group method using an arithmetic average**) (Sokal y Michener, 1958): en este método la distancia entre dos conglomerados se obtiene promediando todas las distancias entre pares de objetos, donde un miembro del par pertenece a uno de los conglomerados y el otro miembro al segundo conglomerado. Este es uno de los métodos más simple y el que se ha encontrado más exitoso en numerosas aplicaciones. Se han propuesto varias expresiones para calcular la distancia promedio, una de ellas es:

$$d_{(AB)C} = \frac{\sum_i \sum_j d_{ij}}{n_{(AB)} n_C}$$

donde d_{ij} es la distancia entre el objeto *i*, que pertenece al conglomerado AB y el objeto *j* que pertenece a al conglomerado C, siendo la sumatoria sobre todos los posibles pares de objetos entre dos conglomerados y donde $n_{(AB)}$ y n_C son los números de objetos en los conglomerados AB y C respectivamente. El método tiende a producir grupos de igual varianza (Milligan, 1980).

Encadenamiento promedio ponderado *(weighted average linkage,* WPGMA*):* conocido también como *método de McQuitty* fue introducido independientemente por Sokal y

Michener (1958) y McQuitty (1966). Representa una generalización del procedimiento anterior usando el número de objetos en cada conglomerado como peso. Vale decir que la distancia se basa en un promedio ponderado. Si los pesos son iguales este método otorga los mismos resultados que el método anterior.

Centroide no ponderado (*unweighted centroid*, UPGMC) (Sokal y Michener, 1958): toma el promedio de todos los objetos en un conglomerado (centroide) para representar al conglomerado y mide las distancias entre objetos o conglomerados con respecto a dicho centroide. Es el procedimiento aglomerativo más robusto a valores extremos (Milligan, 1980).

Centroide ponderado *(weighted centroid):* es un generalización del procedimiento anterior ponderando las distancias por el número de objetos en cada conglomerado que participa en el cálculo como peso. Los métodos basados en el centroide presuponen una matriz de distancia basada en la métrica Euclídea. Si los pesos son iguales este método otorga los mismos resultados que el método anterior.

Ward o método de mínima varianza (Ward, 1963): Es similar al método del centroide, pero cuando une conglomerados realiza una ponderación (por el tamaño de cada grupo) de todos los conglomerados participantes, así en cada unión la pérdida de información es minimizada. Define la distancia entre dos grupos como la suma de las sumas de cuadrados del ANAVA entre los dos grupos sobre todas las variables. El método es recomendado para datos con distribución normal y matrices de covarianzas esféricas, homogéneas entre grupos. Tiende a producir grupos con igual número de observaciones y puede ser muy afectado por valores extremos (Milligan, 1980).

Los procedimientos jerárquicos descriptos anteriormente no realizan ninguna acción diferencial con observaciones aberrantes. Si una observación rara fue clasificada en etapas tempranas del procedimiento en algún grupo, esta permanecerá ahí en la configuración final. Por ello, es importante revisar cuidadosamente las configuraciones finales. La práctica de aplicar más de un procedimiento y más de una medida de distancia, usualmente ayuda a diferenciar entre agrupamientos naturales y artificiales. Algunos experimentadores, usan la técnica de la perturbación (introducción de errores en los datos y reagrupamiento bajo la nueva situación) para probar la estabilidad de la clasificación jerárquica. La técnica de muestreo reiterado conocida como *bootstrap* es también recomendada para probar estabilidad de los nodos logrados en un agrupamiento particular.

Métodos de agrupamiento no jerárquicos

InfoStat permite el agrupamiento de objetos mediante el procedimiento no-jerárquico **K-means**. El algoritmo agrupa objetos en *k* grupos haciendo máxima la variación entre conglomerados y minimizando la variación dentro de cada conglomerado. Este método comienza con un agrupamiento inicial o con un grupo de puntos semilla (centroides) que formarán los centros de los grupos (partición inicial del grupo de objetos en *k* ítems). Prosigue asignando cada objeto al grupo que tiene el centroide (media) más cercano. La distancia comúnmente usada es la Euclídea, tanto en observaciones estandarizadas como en las no estandarizadas. Se trabaja con la minimización de la función objetivo "suma de

distancias al cuadrado". La partición lograda es aquella tal que la suma de la suma de las distancias al cuadrado de los miembros del grupo respecto a su centroide es mínima. El método se basa así en el principio de los *k* mejores centroides. Los centroides son modificados cada vez que un objeto se transfiere de un grupo al otro. El algoritmo K-means es óptimo en cada paso. Los resultados finales podrían depender de la configuración inicial, de la secuencia en que son considerados los objetos a agrupar y claramente del número de grupos. A los fines de alcanzar un óptimo global, es recomendable usar varias particiones iniciales y seleccionar aquella partición final con mínimo valor de la función objetivo. InfoStat reporta automáticamente los valores de dicha función bajo el nombre SSCD (Suma de Sumas de Cuadrado Dentro). La aplicación sucesiva de procedimientos jerárquicos y no-jerárquicos es una estrategia recomendada para determinar el número de grupos más apropiado para el problema en cuestión. Es recomendable aplicar en primera instancia un método jerárquico aglomerativo que sugiera un determinado número de grupos (grupos formados al establecer un criterio de corte como por ejemplo el 55% de la distancia máxima) y luego utilizar dicha información como partición inicial del algoritmo K-means.

Distancias

El análisis de conglomerados requiere medir la similitud entre las entidades a agrupar. InfoStat trabaja con medidas de disimilaridad o distancia. La selección de una medida de distancia apropiada depende de la naturaleza de las variables (cualitativa, cuantitativa), de la escala de medición (nominal, ordinal, intervalo, cociente) y del conocimiento del objeto de estudio. Todas las funciones de distancia discutidas en este documento pueden ser usadas con cualquier procedimiento de formación de conglomerados.

Para datos con propiedades métricas (continuos, escala por intervalos y/o cocientes) pueden usarse medidas de distancia como la de Manhattan o la Euclídea mientras que para datos cualitativos o atributos discretos son más apropiadas las distancias basadas en medidas de similitud o asociación. Distintas funciones pueden ser usadas en InfoStat para obtener distancias a partir de medidas de similitud. Al elegir una medida de similitud en la ventana Análisis de conglomerados, automáticamente se habilita a la derecha de dicha ventana otra ventana donde se puede elegir dicha función.

Para el agrupamiento de variables son recomendadas medidas de distancia basadas en coeficientes de correlación. Todas las medidas de distancia que se pueden usar en este módulo se encuentran descriptas en el módulo del menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ DISTANCIAS ⇒ ASOCIACIONES.

Componentes principales

Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ COMPONENTES PRINCIPALES permite analizar la interdependencia de variables métricas y encontrar una representación gráfica óptima de la variabilidad de los datos de una tabla de *n* observaciones y *p* columnas o variables. El análisis de componentes principales (ACP) trata de encontrar, con pérdida mínima de información, un nuevo conjunto de variables (componentes

principales) no correlacionadas que expliquen la estructura de variación en las filas de la tabla de datos.

En la ventana **Análisis de componentes principales**, se deben indicar las variables respuesta y las de clasificación en caso que existan. En caso de señalar un criterio de clasificación InfoStat trabajará con la matriz *axp* de datos siendo *a* el número de niveles del criterio de clasificación y *p* la cantidad de variables seleccionadas. En la solapa **General** hay opciones para guardar las componentes obtenidas (**Guardar los ejes**) según el número de componentes que se indique o según el criterio utilizado para la selección automática del número de ejes a guardar. Cuando el usuario, activa el casillero **# automático**, InfoStat guardará tantos ejes como autovalores mayores al valor promedio de los autovalores haya. Si se guardan componentes principales se adicionarán como nuevas columnas a la tabla activa. Estas componentes pueden ser utilizadas posteriormente para realizar gráficos de dispersión de las observaciones (el gráfico de dispersión usando como ejes la CP1 y la CP2 permite visualizar la mayor variabilidad entre observaciones). Si se realizan varios ACP, se generarán tantas nuevas columnas como componentes se decidan guardar en cada análisis. Para evitar esta acumulación de nuevas columnas se puede activar la opción **Sobrescribirlos**, así sólo se guardarán las del último ACP. Se puede pedir la estandarización de cada variable antes de comenzar el análisis (**Estandarizar datos)**, la visualización de la matriz de covarianza o correlación (**Mostrar matriz de covarianzas/correlación**) sobre la que se realiza el análisis, la **correlación de cada componente principal con las variables originales**, el coeficiente de **correlación cofenética,** gráficos **Biplot** y Árbol de recorrido mínimo (ARM). En caso de que ese haya indicado un criterio clasificación en la solapa **Medidas de resumen**, InfoStat permite escoger entre medidas de posición como la **media**, **mediana**, **mínimo**, **máximo** y de dispersión como la **varianza** y **desviación estándar** como estadísticos para resumir la información de cada variable en cada conjunto de registros indexado por el criterio (opcional).

Ejemplo 34: En un estudio que tuvo como objetivo estudiar los alimentos que se utilizan como fuentes proteicas, en las dietas de los habitantes de países europeos, se registraron los alimentos consumidos. Los datos se encuentran en el archivo Proteínas.

Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ COMPONENTES PRINCIPALES. En la ventana **Análisis de componentes principales** seleccionar "CarneVacuna", "CarneCerdo" y las demás variables que representan la fuente de proteínas como **Variables** y "País" como **Criterio de clasificación.** En la ventana **Análisis de componentes principales**, se activó **Guardar los ejes** y se introdujo el número 2 para preservar las dos primeras componentes**.** También se activó **Estandarizar los datos** para realizar el análisis sobre la matriz de correlación en lugar de la matriz de covarianza de las variables. Se activó la opción **Biplot** a partir de la cual se obtuvo el siguiente gráfico:

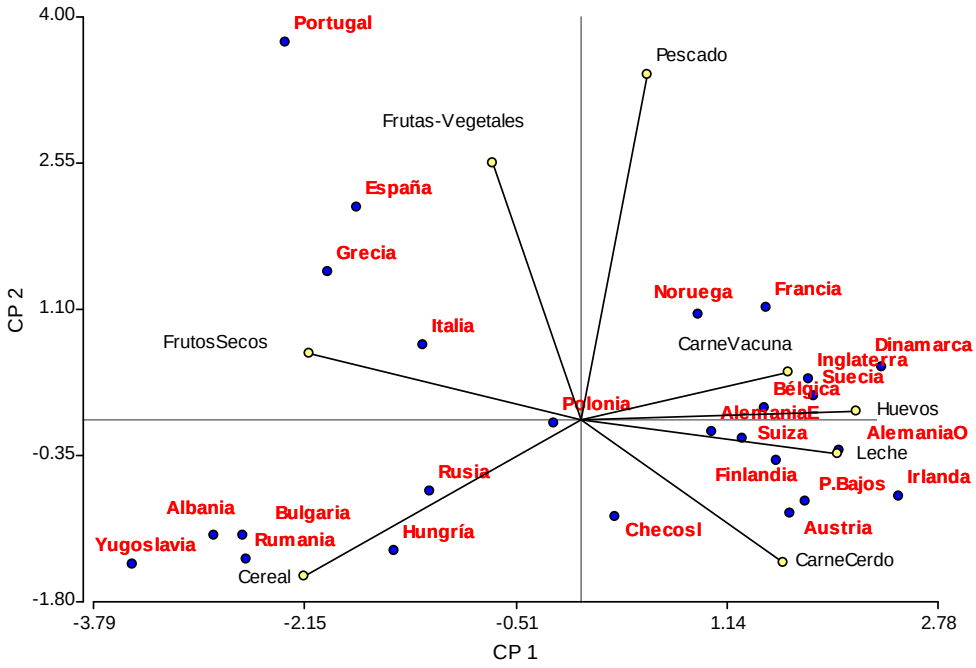


Figura 26: Biplot. Archivo Proteínas.

Como puede verse la primera componente (CP1) separa cereal y frutos secos del resto de las fuentes proteicas, por tanto la mayor variabilidad entre los hábitos de consumo de los distintos países se explica con estas variables. Albania, Bulgaria, Yugoslavia y Rumania están más asociados al consumo de cereal; Grecia e Italia al de frutos secos; España y Portugal consumen más frutas y vegetales; Noruega al consumo de pescado, Alemania del Este, Francia, Dinamarca, Inglaterra, Suecia y Bélgica se asocian al consumo de carne vacuna y en algunos de estos países también huevos. Alemania, Suiza y Finlandia están más asociados al consumo de leche y huevos. Los países Bajos, Irlanda y Austria están asociados a consumo de carne de cerdo y leche, mientras que Checoslovaquia lo esta a la carne de cerdo. Polonia no se encuentra asociado a ninguna fuente proteica particular. Con estos dos ejes se explicó el 65% de la variabilidad total en las observaciones (ver Tabla 52). Para explorar más profundamente estas relaciones se podría haber pedido una tercera componente, solicitando a InfoStat que guarde 3 ejes. En tal caso InfoStat reporta todos los gráficos Biplot que se pueden construir a partir de las tres componentes guardadas.

InfoStat provee automáticamente los autovalores y autovectores resultantes del análisis de Componentes Principales sobre la matriz de correlación (o covarianza según se solicita) de las variables, los cuales se muestran en la ventana **Resultados**. Para el ejemplo presentado dichos valores pueden verse en la Tabla 52. En esta tabla se pueden ver los autovalores asociados a cada autovector (siempre se mostrarán tantos autovectores como componentes principales se hayan seleccionado para el análisis), la proporción de variabilidad total explicada por cada componente (autovalores) y la proporción de la variabilidad total explicada, en forma acumulada, se presenta en la tabla denominada **Autovalores**.

Los resultados de este ejemplo, señalan que con las dos primeras componentes es posible explicar el 65% de la variación total. Los autovectores ($\mathbf{e}_1$ y $\mathbf{e}_2$) reportados muestran los coeficientes con que cada variable original fue ponderada para conformar las CP1 y CP2. En este ejemplo, se puede visualizar que, al construir la CP1, las variables "Cereal" y "FrutosSecos" reciben los pesos negativos más altos y la variable "Huevos" el peso positivo más alto. Las variables "Leche", "CarneVacuna" y "CarneCerdo" también tienen pesos con coeficientes positivos relativamente altos. Luego se puede interpretar que la CP1 opondrá países que utilizan los cereales y frutos secos como principales fuentes proteicas a aquellos que principalmente usan huevos y otros productos de origen animal como fuente proteica. De la misma manera se pueden leer los restantes autovectores retenidos para explicar el significado de cada componente.

En este ejemplo, después de explicar la variabilidad en los hábitos alimentarios de los países debida a la ingesta de cereales y frutos secos versus los productos animal del tipo mencionado, se debería destacar la variabilidad introducida por el consumo o no de pescado y frutas y vegetales (CP2). La ortogonalidad de las componentes principales garantiza que la CP2 provee nueva información sobre variabilidad respecto a la provista por la CP1, es decir explica variabilidad en los hábitos alimenticios entre países no explicada por la CP1.

Tabla 52: Análisis de componentes principales: autovalores y autovectores. Archivo Proteínas.

Análisis de componentes principales-Datos estandarizados

Autovalores

Lambda	Valor	Proporción	Prop Acum
1	3.72	0.47	0.47
2	1.50	0.19	0.65
3	1.09	0.14	0.79
4	0.85	0.11	0.90
5	0.33	0.04	0.94
6	0.28	0.03	0.97
7	0.12	0.02	0.99
8	0.10	0.01	1.00

Autovectores

Variables	e1	e2
CarneVacuna	0.33	0.10
CarneCerdo	0.33	-0.29
Huevos	0.44	0.02
Leche	0.41	-0.07
Pescado	0.11	0.71
Cereal	-0.44	-0.32
FrutosSecos	-0.44	0.14
Frutas-Vegetales	-0.14	0.53

Las componentes principales generadas por InfoStat resultan de la combinación lineal de las variables originales previamente estandarizadas cuando el análisis se aplica a la matriz de correlación. Cuando el análisis se aplica a la matriz de covarianzas, InfoStat reporta combinaciones lineales de variables sólo centradas por su media.

Si además de solicitar el gráfico Biplot, en el ejemplo anterior se activaba la opción **ARM** se obtiene el siguiente gráfico:

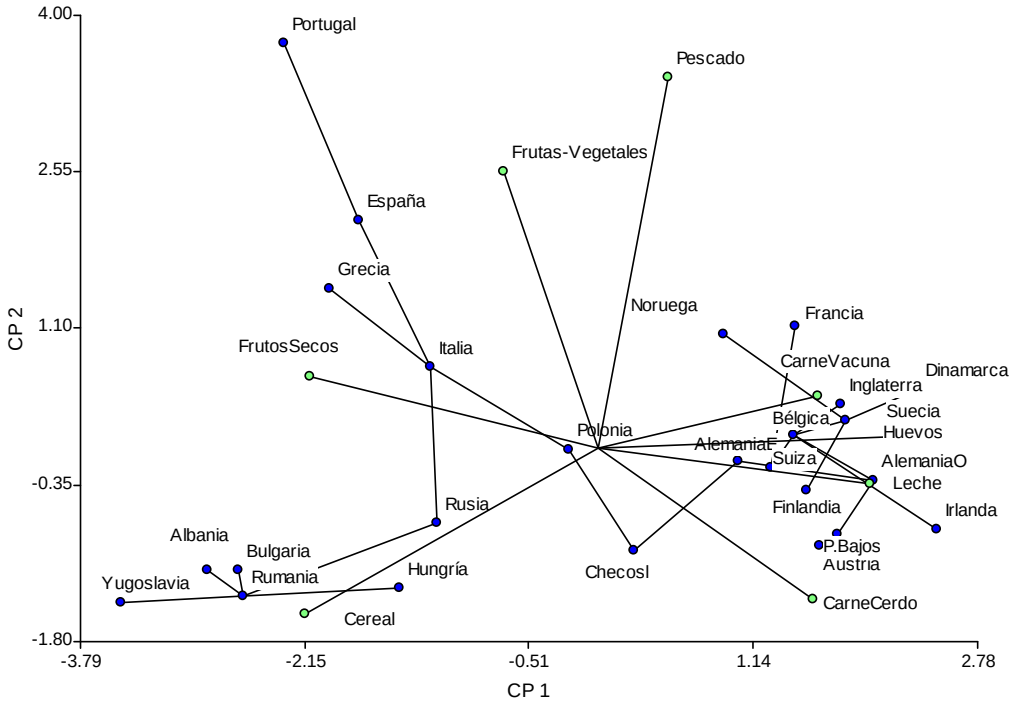


Figura 27: Biplot más árbol de recorrido mínimo. Archivo Proteínas.

Los árboles de recorrido mínimo (ver **ARM**) unen los puntos observación de acuerdo a la distancia entre ellos calculada en el espacio original, es decir aquel de tantas dimensiones como variables participen en el estudio. La distancia entre dos puntos en el plano (espacio de reducción reducida respecto al original), puede no reflejar fielmente la estructura de distancias verdaderas, es decir aquella del espacio de mayor dimensión. En este ejemplo, la opción **ARM** permite visualizar mejor las asociaciones entre países en función de su fuente de proteínas. Por ejemplo, Hungría esta más cerca en el grafico de Rusia que de Rumania, sin embargo, los hábitos alimentarios de Hungría se parecen mas a los de Rumania que a los de Rusia ya que el árbol los une primero.

Nociones teóricas sobre el análisis de componentes principales

El análisis de componentes principales y los gráficos conocidos como *biplot* son técnicas generalmente utilizadas para reducción de dimensión. Las técnicas de reducción de dimensión permiten examinar todos los datos en un espacio de menor dimensión que el espacio original de las variables. Con el ACP se construyen ejes artificiales *(componentes principales)* que permiten obtener gráficos de dispersión de observaciones y/o variables con propiedades óptimas para la interpretación de la variabilidad y covariabilidad subyacente. Los biplots permiten visualizar observaciones y variables en un mismo espacio, así es posible identificar asociaciones entre observaciones, entre variables y entre variables y observaciones.

Diferencias en los datos generan variabilidad, luego una forma de resumir y ordenar los datos es a través del análisis o la explicación de la estructura de varianza y covarianza del conjunto de variables en estudio. El ACP es una técnica frecuentemente utilizada para ordenar y representar datos multivariados continuos a través de un conjunto de *d=1,...,p combinaciones lineales* ortogonales normalizadas de las variables originales que explican la variabilidad existente en los datos de forma tal que ningún otro conjunto de combinaciones lineales de igual cardinalidad, tiene varianza de las combinaciones mayor a la del conjunto de componentes principales. Usualmente se selecciona un número *d* mucho menor que *p*, para la representación de la variabilidad subyacente. Se espera que dicha reducción de dimensionalidad no produzca una pérdida importante de información. Desde este punto de vista, la técnica de reducción de la dimensión implica una consecuente ayuda en la interpretación de los datos. La primera componente contiene más información (sobre variabilidad) que la segunda, ésta a su vez más que la tercera y así sucesivamente hasta no explicar más variabilidad.

El ACP para ordenar observaciones se basa en la descomposición espectral de la *matriz de covarianzas* o de *correlación* entre variables de dimensión *p*x*p*. La selección entre el *estimador insesgado* y el *estimador máximo-verosímil* de la matriz de covarianza poblacional es irrelevante, ya que produce las mismas componentes principales muestrales. Usando los autovectores de **S** o **R** como vectores de coeficientes para la combinación lineal se puede demostrar que las componentes principales son combinaciones lineales no correlacionadas cuyas varianzas son máximas.

La j-ésima componente principal (CPj) es algebraicamente una combinación lineal de las *p* variables originales obtenida como $Y_j = \mathbf{e}_j{}'\mathbf{X} = e_{1j}X_1 + e_{2j}X_2 + ... + e_{pj}X_p$ con j=1,...,p donde $\mathbf{e}_j$ representa el j-ésimo autovector. Las nuevas variables, CP, usan información contenida en cada una de las variables originales, algunas variables pueden contribuir más a la combinación lineal que otras. La varianza de la j-ésima componente principal es $Var(Y_j) = e_j{}'\Sigma e_j = \lambda_j$ donde el j-ésimo λ es el autovalor asociado al j-ésimo autovector de **S**, (los autovalores se ordenan en forma decreciente, $\lambda_1>\lambda_2>\lambda_3$....). Además se satisface que entre dos componentes cualesquiera, la covarianza es nula. La proporción de la varianza total explicada por las primeras *d* componentes será:

$$\text{Prop}_d = \frac{\lambda_1 + \lambda_2 + ... + \lambda_d}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

Los coeficientes de cada variable original estandarizados para una CP, permiten identificar las variables con mayor contribución en la explicación de la variabilidad entre observaciones en el eje asociado a la CP correspondiente. Para analizar asociación entre variables con componentes se pueden solicitar las correlaciones entre las componentes principales y las variables originales. Estas vienen dada por:

$$r(Y_j, X_k) = \frac{e_{kj}\sqrt{\lambda_j}}{\sqrt{\sigma_k^2}}$$

y representan también un indicador de cuán importante es una variable particular en la construcción de la componente. La interpretación de esta correlación puede ser más confiable que la interpretación de los coeficientes que conforman los autovectores, ya que la correlación tiene en cuenta diferencias en las varianzas de las variables originales y consecuentemente elimina el sesgo de interpretaciones causadas por diferentes escalas de medición.

Los datos a analizar podrían o no ser previamente centrados y/o escalados dando lugar a distintos tipos de ACP. El ACP a partir de la matriz de correlación (matriz de covarianzas de las variables originales centradas y escaladas) de los datos es útil cuando las unidades de medida y/o las varianzas de las variables son diferentes. De otro modo las variables con mayor varianza (no necesariamente más informativas) tendrán demasiada influencia en la determinación de la solución. Las componentes principales obtenidas usando la matriz de correlación pueden ser sustancialmente diferentes a las obtenidas usando la matriz de covarianza. En cada caso habrá que juzgar el análisis más conveniente. Cuando las variables no tienen varianzas similares o no son medidas sobre la misma escala (variables no conmensurables), se recomienda la obtención de las componentes a partir de la matriz de correlación.

El coeficiente de correlación cofenética en el marco del ACP reportado por InfoStat, calcula la correlación entre las distancias Euclídeas en el espacio reducido y las mismas distancias en el espacio original de dimensión dada por el número de variables originales. Luego, este valor puede ser utilizado como una medida de la calidad de la reducción lograda.

Biplot

Los gráficos de dispersión construidos a partir de las Componentes Principales pueden ser usados para visualizar la dispersión de las observaciones pero la influencia de las variables no es explícita en tales diagramas. Los gráficos biplots propuestos por Gabriel (Gabriel, 1971), muestran las observaciones y las variables en el mismo gráfico, de forma tal que se pueden hacer interpretaciones sobre las relaciones conjuntas entre observaciones y variables. El prefijo "bi" en el nombre biplot refleja esta característica, tanto observaciones como variables son representadas en el mismo gráfico.

En los biplots, InfoStat grafica las observaciones como puntos azules. La configuración de los puntos es obtenida a partir de un ACP. Las variables son graficadas como vectores desde el origen (con terminaciones en círculos amarillos). En los biplots construidos, la distancia entre símbolos representando observaciones y símbolos representando variables no tiene interpretación, pero las direcciones de los símbolos desde el origen sí pueden ser interpretadas. Las observaciones (puntos filas) que se grafican en una misma dirección que una variable (punto columna) podría tener valores relativamente altos para esa variable y valores bajos en variables o puntos columnas que se grafican en dirección opuesta. Por otro lado, los ángulos entre los vectores que representan las variables pueden ser interpretados en términos de las correlaciones entre variables. Ángulos de 90° entre dos variables indican que ambas variables no se encuentran correlacionadas. Alejamientos de este valor (tanto sea en valores menores como mayores a 90°) implican correlación (positiva o negativa). Es decir un ángulo cercano a cero implica que ambas variables están fuertemente correlacionadas en

forma positiva y un ángulo cercano al ángulo llano entre dos variables indica que ambas muestran fuerte correlación negativa. Cuando las longitudes de los vectores son similares el gráfico sugiere contribuciones similares de cada variable en la representación realizada.

En InfoStat, la opción **Biplot** puede ser seleccionada para representar los resultados del ACP. El gráfico se obtiene realizando un diagrama de dispersión de las observaciones a partir del ACP sobre la matriz de correlación (covarianzas) de las variables y superponiendo los autovectores que representan las variables convenientemente escaladas en el mismo espacio.

Arboles de Recorrido Mínimo (ARM)

Los árboles de recorrido se construyen uniendo puntos que representan observaciones multivariadas y que se proyectan en un plano como resultado de alguna técnica de reducción de dimensión. Los puntos son conectados con segmentos de líneas rectas tal que todos los puntos quedan unidos directa o indirectamente y no hay loops (Gower y Ross, 1969). El árbol de mínimo recorrido es un árbol de recorrido con segmentos conectados de tal manera que la suma de las longitudes de todos los segmentos es mínima. Dentro del ACP los ARM se calculan teniendo en cuenta la distancia de los puntos filas de la matriz de datos en el espacio original (cuya dimensión es igual al número de variables en estudio).

Análisis discriminante

Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DISCRIMINANTE, permite realizar el análisis multivariado discriminante (AD) canónico para variables métricas. El análisis discriminante es útil para: 1) **discriminar**, en base a las variables seleccionadas grupos definidos *a priori* y poder representar a las observaciones en un espacio donde las diferencias entre grupos sean máximas y 2) **clasificar** nuevos casos en los grupos establecidos *a priori* sobre la base de una regla de clasificación basada en las variables independientes.

En la ventana **Análisis discriminante lineal**, declarar las **Variables** que conformarán la función discriminante y la o las variables que definen los grupos en **Criterio de Agrupamiento**. InfoStat espera que exista al menos más de una observación por grupo. Para que el análisis pueda realizarse, el número de individuos por grupo debería ser superior al número de variables o por lo menos igual. Al **Aceptar** aparece otra ventana donde se puede solicitar: la **Matriz de covarianzas entre grupos**, la **Matriz de sumas de cuadrados entre grupos**, la **Matriz de covarianzas residual (Común)**, la **Matriz de sumas de cuadrados residual (Común)**, los **Análisis de la varianza univariados**, **Mostrar tasas de error de clasificación**, **Guardar primeras coordenadas discriminantes** (al lado de esta opción se puede elegir el número de coordenadas a guardar), **Sobreescribir coordenadas discriminantes, Gráfico** de las dos primeras coordenadas discriminantes, gráficos **Biplot** y **ARM** o árboles de recorrido mínimo.

Las opciones **Matriz de covarianzas entre grupos** y **Matriz de covarianzas residual (Común)** permiten visualizar las matrices de covarianzas relacionadas a la hipótesis de

efectos de grupos (**H**) y a las varianzas y covarianzas residuales obtenidas después de descontar el efecto de grupos (**E**). También se pueden requerir las matrices de sumas de cuadrados que originan las matrices **H** y **E**. InfoStat permite obtener los **análisis univariados de la varianza**, *i.e*. análisis basados en el estadístico F construido a partir de los elementos diagonales de las matrices **H** y **E** para cada variable.

Activando **Mostrar tasas de error de clasificación**, InfoStat proporciona las tasas de error aparente (estimadores de la probabilidad de una mala clasificación) obtenidas al clasificar las observaciones del archivo en los grupos en cuestión a partir del uso de la función discriminante recientemente construida. Es común en problemas de este tipo, cuando se tienen suficientes datos, particionar el conjunto de datos en dos subconjuntos, uno es utilizado para hallar la función discriminante y otro para la validación de la misma. La función estimada a partir del primer archivo (*datos de calibración*) puede ser evaluada con los datos del segundo archivo (*datos de validación*). Las tasas de error aparente utilizan un único archivo para ambos procesos, es decir las mismas observaciones usadas para estimar la función son luego reclasificadas con la función para estimar el error de clasificación. Las tasas de error aparente tienden a subestimar el error, son útiles cuando se disponen de grandes tamaño de muestra en cada población (Johnson y Wichern, 1998). Seleccionando **Guardar coordenadas discriminante** se genera en la tabla de datos tantas nuevas columnas como se indique (el máximo número a especificar es igual al número de grupos menos 1 o el número de variables según cual sea menor).

InfoStat realiza automáticamente la prueba de Barttlet para la hipótesis de **homogeneidad de matrices de covarianza** (Morrison, 1976), tal prueba es útil para definir si es óptimo trabajar con una función discriminante lineal o si se debiera usar otro tipo de función discriminante. El rechazo de la hipótesis nula de la prueba (hipótesis de igualdad de matrices de covarianza para cada grupo) sugiere, cuando los datos son normales, que una función de discriminación cuadrática sería más adecuada que la función discriminante lineal. También se reporta automáticamente el número máximo de **funciones discriminante canónicas** (excepto cuando se requiere guardar un número menor), las **funciones discriminantes estandarizadas por la matriz de covarianza residual común** y los **centroides en el espacio discriminante**, es decir los valores promedios de las funciones canónicas para cada grupo. Estos centroides son útiles para la clasificación posterior de nuevas observaciones en los grupos. La construcción de las funciones canónicas se realiza de acuerdo a las nociones teóricas explicitadas más adelante.

Ejemplo 35: Se utilizan los datos del archivo Iris conteniendo 50 observaciones de 4 características de una flor: longitud del sépalo, ("SepalLen"), ancho del sepalo ("SepalWid"), longitud del pétalo ("PetalLen") y ancho del pétalo ("PetalWid") para 3 especies del género Iris (Fisher, 1936). Total de observaciones: 150. Se pretende encontrar una función discriminante que permita clasificar nuevas flores en uno de los tres grupos (especies), de acuerdo al valor que asumen para esas flores, las cuatro variables que conforman la función discriminante.

Después de seleccionar el AD, en la ventana **Análisis Discriminante Lineal**, se declaró "SepalLen", "SepalWid", "PetalLen" y "PetalWid" como **Variables** y Especie como **Criterio de agrupamiento**. Al **Aceptar** aparece otra ventana de diálogo donde se

solicitaron las opciones por defecto: **Mostrar tasas de error de clasificación**, **Guardar coordenadas discriminantes** (2, igual al número de grupos menos 1) y **Sobreescribir coordenadas discriminantes** (si en el archivo ya existe una columna conteniendo los ejes canónicos (coordenadas discriminantes) 1 y 2, InfoStat sobrescribirá dichas columnas con los ejes obtenidos en el análisis presente). Luego se seleccionó **Grafico.** Al **Aceptar** se obtuvieron los siguientes resultados:

Tabla 53: Análisis discriminante lineal. Archivo Iris.

Análisis discriminante lineal

Prueba de Homogeneidad de Matrices de Covarianzas

Grupos	N	Estadístico	gl	p-valor
3	150	140.94	20	<0.0001

Autovalores de Inv(E)H

Autovalores	%	% acumulado
32.19	99.12	99.12
0.29	0.88	100.00

Funciones discriminantes canónicas

	1	2
Constante	-2.11	-6.66
SepalLen	-0.83	0.02
SepalWid	-1.53	2.16
PetalLen	2.20	-0.93
PetalWid	2.81	2.84

Funciones discriminantes-datos estandarizadas con la varianzas comunes

	1	2
SepalLen	-0.43	0.01
SepalWid	-0.52	0.74
PetalLen	0.95	-0.40
PetalWid	0.58	0.58

Centroides en el espacio discriminante

Grupo	Eje 1	Eje 2
1	-7.61	0.22
2	1.83	-0.73
3	5.78	0.51

Tabla de clasificación cruzada

Grupo	1	2	3	Total	Error(%)
1	50	0	0	50	0.00
2	0	48	2	50	4.00
3	0	1	49	50	2.00
Total	50	49	51	150	2.00

La prueba de homogeneidad de matrices de covarianzas arrojó un valor $p<0.001$, sugiriendo que no se cumple este supuesto y que una función discriminante cuadrática podría ser mejor. Sin embargo se continuó con el análisis ya que este conjunto de datos ha sido ampliamente utilizado en la literatura para ejemplificar los resultados del AD lineal. A partir de los autovalores de la expresión inv(**E**)**H**, se puede concluir que el eje canónico 1 explica el 99.12% de la variación entre grupos. Como hay tres grupos se generaron dos funciones discriminantes, o sea dos ejes canónicos, el valor de cada observación sobre los cada eje canónico se adicionan a la tabla de datos. La primera función **discriminante canónica** puede expresarse de la siguiente manera:

F=-2.11-0.83(SepalLen)-1.53(SepalWid)+2.20(PetalLen)+2.81(PetalWid)

En esta función lineal de las cuatro variables seleccionadas, los coeficientes responden a las distribuciones de cada variable. Si las variables tienen varianzas muy distintas y/o existe alta covariación entre pares de variables, la interpretación puede ser engañosa, por eso conviene analizar la importancia relativa de cada variable, en la discriminación de los grupos, usando la función con coeficientes estandarizados por varianzas y covarianzas. A partir de la primera **función discriminante estandarizada por las covarianzas comunes** puede verse que PetalLen es la variable más importante para la discriminación sobre este eje. Observaciones (flores) con valores altos para esta variable (pétalos más largos) aparecerán situadas a la derecha del gráfico de dispersión de observaciones en el espacio discriminante (espacio formado por los ejes canónicos) ya que el coeficiente es positivo (0.95).

Los **centroides en el espacio discriminante** o medias de las funciones por grupo, muestran que el Grupo 1 se opone a los otros dos grupos en el eje canónico 1, indicando que las diferencias en PetalLen permiten discriminar observaciones del grupo 1 (menor longitud de pétalos) respecto a aquellas de los grupos 2 y 3. De igual manera se pueden interpretar diferencias entre grupos usando el eje canónico 2. En este ejemplo el eje 2 explica muy poca variación entre los grupos (el autovalor asociado señala que el porcentaje de variación explicada sobre este eje es 0.88%). Por ello se debe señalar la importancia relativa de los ejes canónicos.

La **tabla de clasificación cruzada** que se presenta al final de la salida (en filas se representa el grupo al que pertenece la observación y en columnas el grupo al que es asignada la misma observación al usar la función discriminante) señala que las 50 plantas del Grupo 1 fueron todas bien clasificadas, la tasa de error de clasificación en este grupo es 0%. De los 50 individuos del Grupo 2, 48 fueron asignados bien y dos fueron mal clasificados dentro del Grupo 3, la tasa de error es del 4%. Similar interpretación se hace para el Grupo 3. La tasa de error aparente promedio es del 2%. InfoStat adiciona automáticamente a la tabla de datos una columna que se llama "Clasificación", donde puede verse que los casos 71, 84 y 134 fueron aquellos mal clasificados.

Para visualizar la discriminación entre grupos sugerida por el AD, se seleccionó Grafico en la ventana de AD. Esta opción produce automáticamente un diagrama de dispersión con el eje canónico 1 y el eje canónico 2, particionado por el criterio de clasificación, en este caso "especie". Al gráfico se le agregaron las **elipses de predicción**, las que se logran de la siguiente manera: seleccionar las tres series, presionar el botón derecho y elegir "dibujar contornos", esto habilita el submenú "opciones del contorno", que son: "contorno simple", "elipse de predicción" y "elipse de confianza". Se marcaron además las tres observaciones erróneamente clasificadas.

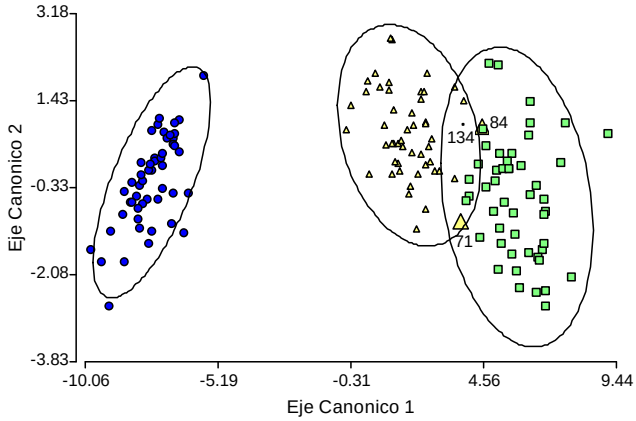


Figura 28: Representación de observaciones multivariadas en tres grupos, definidos a priori, en el espacio discriminante conformado por los ejes canónicos 1 y 2 del AD. Contornos corresponden a elipses de predicción. Archivo Iris.

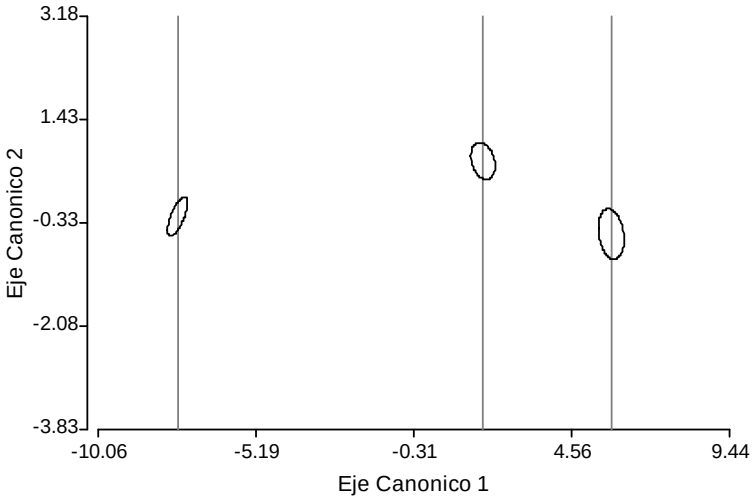


Figura 29: elipses de confianza para el centroide de observaciones multivariadas de tres grupos en el espacio discriminante. Líneas verticales corresponden al centroide de cada grupo sobre el eje canónico 1. Archivo Iris.

Nociones teóricas sobre análisis discriminante

El AD permite describir algebraicamente las relaciones entre dos o más poblaciones (grupos) de manera tal que las diferencias entre ellas se maximicen o se hagan más evidente. El AD se realiza frecuentemente con fines predictivos relacionados a la clasificación, en una de las poblaciones existentes, de nuevas observaciones u observaciones sobre las cuales no se conoce a qué grupo pertenecen. Una observación *nueva*, la cual no fue utilizada para la

construcción de la regla de clasificación, se asignará al grupo en el cual tienen más probabilidad de pertenecer en base a sus características medidas. Para tal asignación es necesario definir una regla de clasificación. La función discriminante lineal puede ser usada para este fin. Además el AD puede ser usado con el objetivo de encontrar el subconjunto de variables que mejor explica la variabilidad entre grupos.

El análisis presupone que se dispone de *n* observaciones p-dimensionales independientes, las cuales se encuentran agrupadas en dos o más grupos. El AD abordado por InfoStat presupone que la variable dependiente es nominal (forma grupos) y las variables independientes son métricas (variables continuas, escala por intervalos o cociente). Es decir la variable que actúa como factor de agrupamiento y ubica cada individuo u objeto de la tabla de datos en un grupo definido. Este agrupamiento es conocido *a priori* de realizar el análisis. Por ejemplo si se está estudiando el grado de ataque de una virosis, que afecta los pobladores de un área, y se determinan tres grados de ataque tal que 0 denota "bajo", 1 "medio" y 2 "alto", cada individuo de una muestra de pobladores debe pertenecer sólo a uno de estos grupos. Además se supone que existe una cantidad *p* de variables con potencialidad para explicar la clasificación o agrupamiento de los pobladores de acuerdo al grado de ataque experimentado.

El análisis discriminante lineal puede interpretarse en analogía con el análisis de regresión lineal múltiple univariada. El objetivo del análisis de regresión es predecir el valor de la media poblacional de una variable dependiente, sobre la base de una combinación lineal de variables independientes para las cuales se conocen los valores que asumen en los individuos de una muestra. El objetivo del análisis discriminante es encontrar una combinación lineal de variables independientes que minimice la probabilidad de clasificar erróneamente individuos u objetos en sus respectivos grupos. En cuanto a los supuestos, en el análisis de regresión la variable dependiente se asume normalmente distribuida y las independientes son fijas. En el análisis discriminante las variables independientes son las que generalmente se asumen normalmente distribuidas y la dependiente (variable de agrupamiento) es fija.

La función discriminante calculada por InfoStat es una combinación lineal de las variables originales, en la que la suma de cuadrados de las diferencias entre grupos, para dicha combinación, sobre la varianza dentro de los grupos es máxima. Cuando hay dos grupos se genera una sola ecuación lineal discriminante (eje canónico). Si hay *k* grupos, habrá *k*-1 funciones discriminantes no correlacionadas (ejes canónicos). Se recomienda graficar las observaciones en el espacio generado por los ejes canónicos para obtener una mejor visualización de las diferencias entre grupos.

Cuando los datos asumen una distribución normal multivariada el análisis puede realizarse por métodos paramétricos ya sea asumiendo homogeneidad en la estructura de variación y covariación y por ende trabajando con la matriz de covarianzas común (en tal caso la función discriminante será lineal) o trabajando a partir de las matrices de covarianza de cada grupo (función discriminante cuadrática). La violación del supuesto de normalidad se puede solucionar transformando las variables hasta conseguir normalidad o realizando un AD no paramétrico. Estudios de simulación muestran que la función discriminante lineal es muy robusta a alejamientos del supuesto de normalidad multivariada. La falta de cumplimiento

del supuesto de homogeneidad de matrices de covarianza cuando se usa una función lineal puede aumentar el error en la clasificación.

La primera aproximación al problema de discriminación lineal para *k*=2 grupos fue sugerida por Fisher (1936) quien abordó el problema desde una óptica univariada usando una combinación lineal de las características observadas. Sea $\mathbf{x}$ el vector *p*x1 de características medidas sobre un elemento de una población y consideremos dos poblaciones π_1 y π_2. Llamaremos $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ a las funciones de densidad multivariadas asociadas con las poblaciones π_1 y π_2. respectivamente. Asumimos que las variables aleatorias caracterizadas por estas funciones multivariadas tienen vectores medios $\mu_1 = E(\mathbf{x} \mid \pi_1)$ y $\mu_2 = E(\mathbf{x} \mid \pi_2)$ y matriz de covarianza común $\Sigma_1 = \Sigma_2 = \Sigma = E(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)'$ i=1,2.

Consideremos la combinación lineal $\mathbf{y} = l'\mathbf{x}$, luego se tiene que $\mu_{1y} = E(\mathbf{y} \mid \pi_1) = l'\mu_1$ y $\mu_{2y} = E(\mathbf{y} \mid \pi_2) = l'\mu_2$ y que la varianza de la combinación lineal es $V(\mathbf{y}) = \sigma_y^2 = l'\Sigma l$.

La idea de Fisher fue maximizar la distancia estadística entre μ_{1y} y μ_{2y} a través de una selección apropiada del vector de coeficientes de la combinación lineal, es decir:

$$\text{maximizar}\left(\frac{(\mu_{1y} - \mu_{2y})^2}{\sigma_y^2}\right) = \text{maximizar}\left(\frac{(l'\mu_1 - l'\mu_2)^2}{l'\Sigma l}\right)$$

La solución a dicha maximización es $l = c\Sigma^{-1}(\mu_1 - \mu_2) \;\forall\; c \neq 0$. La combinación lineal del vector de observaciones y el vector *l* es conocida como función lineal discriminante de Fisher, $\mathbf{y} = l'\mathbf{x} = (\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x}$.

Para clasificar una observación nueva, $\mathbf{x}_0$, usando la función lineal discriminante de Fisher obtendremos el "score" para $\mathbf{x}_0$ calculando $\mathbf{y}_0 = l'\mathbf{x}_0 = (\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x}_0$.

Si se define $m = \frac{1}{2}(\mu_{1y} + \mu_{2y}) = \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$ como el punto medio entre las medias univariadas de $\mathbf{y}$, tenemos que:

$$E(\mathbf{y}_0 \mid \pi_1) - m \geq 0 \quad \text{y} \quad E(\mathbf{y}_0 \mid \pi_2) - m < 0.$$

Luego, la regla de clasificación será asignar $\mathbf{x}_0$ a la población π_1 si $\mathbf{y}_0 \geq m$ y a la población π_2 si $\mathbf{y}_0 < m$. Es decir si la nueva observación se encuentra más cercana al centroide del grupo 1 en el espacio discriminante será asignada al grupo 1, en caso contrario se clasificará en el grupo 2.

Cuando más de dos grupos o poblaciones describen la estructura de las observaciones, el método de Fisher es generalizado bajo el nombre de análisis discriminante canónico. Sea

$\mathbf{H} = \sum_{i=1}^{g} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$ donde $\overline{\mathbf{x}} = \frac{1}{g}\sum_{i=1}^{g} \overline{\mathbf{x}}_i$ la matriz de sumas de cuadrados y productos cruzados entre grupos o matriz de SCPC asociada a la hipótesis **H** sobre efectos de grupos y definamos a la matriz común de SCPC de los términos de error como $\mathbf{E} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)' = \sum_{i=1}^{g}(n_i - 1)\mathbf{S}_i$. Los autovectores de $\mathbf{E}^{-1}\mathbf{H}$ son las funciones discriminantes canónicas que separan los grupos. La regla de clasificación en este caso sugiere asignar $\mathbf{x}_0$ en el grupo con media más cercana, en términos de distancia estadística, a $\mathbf{x}_0$.

Luego, $\mathbf{x}_0$ se deberá asignar a π_k si $\sum_{j=1}^{r}[l_j'(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_k)]^2 \leq \sum_{j=1}^{r}[l_j'(\overline{\mathbf{x}}_0 - \overline{\mathbf{x}}_i)]^2$ para todo $i \neq k$ *y* $r \leq s = \min(g-1, p)$.

El primer eje canónico (asociado con el mayor de los autovalores, λ_i, de $\mathbf{E}^{-1}\mathbf{H}$) permite visualizar la máxima separación entre los grupos. En la práctica, sólo los primeros ejes pueden ser necesarios para explicar la separación entre los grupos. Los autovalores dan una medida de la separación entre grupos en la dirección dada por el autovector (coeficientes de la combinación lineal) correspondiente. Si s es el máximo número de ejes canónicos que se pueden obtener, el valor $\frac{\lambda_1 + ... + \lambda_r}{\lambda_1 + ... + \lambda_s}$ es la proporción de variación entre grupos explicada por los primeros *r* ejes canónicos. Cuando sólo dos o tres variables canónicas describen apropiadamente la separación entre grupos, generalmente se grafican las observaciones en el espacio definido por esos ejes para reducir la dimensión de la representación. Luego, se puede pensar al AD como una técnica de reducción de dimensión relacionada al ACP. En el AD lineal se obtienen variables o ejes canónicos a partir de la combinación lineal de variables cuantitativas, tal que dicha combinación explica la variación entre clases o grupos de la misma forma que las combinaciones lineales que constituyen componentes principales en el ACP explican la variación total.

Una forma útil de obtener una medida de la importancia de una variable respuesta sobre una variable canónica (o eje canónico) es a través de la estandarización de los coeficientes de la combinación lineal correspondiente. Si $D = [Diag(\mathbf{E})]^{1/2}$ es la matriz diagonal de desviaciones estándar de las variables originales, luego $l_s = \mathbf{D}\, l$ es el vector de coeficientes estandarizados derivados desde los coeficientes canónicos en ***l***. Estos coeficientes son útiles para juzgar la contribución de cada variable original en la explicación de la variabilidad entre grupos.

Correlaciones canónicas

Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ CORRELACIONES CANÓNICAS, permite calcular correlaciones canónicas (correlaciones entre grupos de variables) y probar su significancia estadística.

InfoStat produce automáticamente una serie de pruebas de hipótesis que establecen que cada correlación canónica y todas las menores son cero en la población (Johnson y Wichern, 1998). InfoStat usa la aproximación usual basada en el estadístico Chi cuadrado, es importante que al menos uno de los dos conjuntos tenga una distribución aproximadamente normal multivariada para que los niveles de probabilidad sean válidos. En la salida se podrán observar, para cada una de las correlaciones canónicas factibles de calcular, el coeficiente de correlación canónica (R), la proporción de la varianza total explicada por cada par de variables canónicas (R^2), el estadístico (lambda), para probar la hipótesis de que dicha correlación y todas las menores son iguales a cero en la población, los grados de libertad (gl) y los niveles de probabilidad asociados a dicha prueba (valor *p*).

Los coeficientes asociados a cada variable original en la generación de una variable canónica pueden ser estandarizados o no estandarizados. InfoStat también realiza **correlaciones canónicas parciales** si se especifica un criterio de agrupamiento a nivel de las observaciones (no de las variables). El análisis de correlaciones canónicas parciales (Timm, 1975) es una generalización multivariada del análisis común de correlación parcial y se interpreta de igual manera que el análisis de correlaciones canónicas (ACC). Las correlaciones son obtenidas a partir de la matriz de varianzas y covarianzas residual obtenida después de ajustar por efectos de grupos.

InfoStat adiciona automáticamente a la tabla de datos los valores que asumen cada una de las variables canónicas (*score* de cada observación sobre cada combinación lineal definiendo una variable canónica). Las correlaciones entre las variables originales y las variables canónicas pueden solicitarse desde el menú ANÁLISIS DE CORRELACIÓN.

El análisis de regresión lineal múltiple entre una variable canónica y todas las variables originales en el conjunto opuesto, puede ser realizado en el menú REGRESIÓN LINEAL para facilitar la interpretación de la correlación canónica. Gráficos de dispersión de cada variable canónica versus su contraparte en el otro grupo son también recomendados.

Para realizar un ACC en InfoStat, en **Variables** de la ventana **Correlaciones canónicas** se deben señalar las variables que conforman el primer grupo (variables en el grupo 1 o variables dependientes) y las que conforman el segundo grupo (variables en el grupo 2 o variables independientes). Cuando se **Acepta,** aparece otra ventana en la cual se puede elegir utilizar las **Variables en su escala original (usa matriz de covarianzas)** o **Variables estandarizadas (usa matriz de correlación)**.

Ejemplo 36: en un estudio realizado con alumnos del último año de la escuela secundaria, se deseaba conocer si las calificaciones en asignaturas de naturaleza cuantitativa como Matemática, Física y Contabilidad se correlacionaban o no con las calificaciones obtenidas en asignaturas de naturaleza no cuantitativa como Lengua, Literatura e Historia. El estudio se realizó analizando los resultados de 6 evaluaciones, una para cada asignatura, por estudiante sobre una muestra aleatoria de alumnos. Los docentes responsables de este

ensayo opinaban que los alumnos que tenían buen desempeño en las materias de naturaleza cuantitativa lo tendrían también en las materias no cuantitativas. Los datos se encuentran en el archivo CorrCan.

En el Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ CORRELACIONES CANÓNICAS, ventana **Correlaciones canónicas**, se seleccionaron en **Variables**, Matemática, Física y Contabilidad para el grupo 1 y Lengua, Literatura e Historia para el grupo 2. En la ventana siguiente se eligió utilizar: **Variables estandarizadas (usa matriz de correlación)**. En la ventana de **Resultados** se generó la salida siguiente:

Tabla 54: Correlaciones canónicas. Archivo CorrCan.

Matriz de correlación

	Literatura	Historia	Lengua	Matematica	Fisica	Contabilidad
Literatura	1.000	0.597	0.853	0.870	0.127	0.865
Historia	0.597	1.000	0.778	0.768	0.226	0.566
Lengua	0.853	0.778	1.000	0.982	0.166	0.760
Matemática	0.870	0.768	0.982	1.000	0.134	0.738
Física	0.127	0.226	0.166	0.134	1.000	0.347
Contabilidad	0.865	0.566	0.760	0.738	0.347	1.000

Correlaciones canónicas

	L(1)	L(2)	L(3)
R	0.990	0.601	0.148
R^2	0.980	0.361	0.022
Lambda	68.246	7.297	0.344
gl	9.000	4.000	1.000
p-valor	0.000	0.121	0.558

Coef. combinaciones lineales

	L(1)	L(2)	L(3)
Literatura	0.271	1.879	-0.470
Historia	0.036	-0.066	-1.624
Lengua	0.731	-1.687	1.692
Matemática	0.845	1.223	0.261
Física	-0.018	0.478	-0.976
Contabilidad	0.202	-1.578	-0.118

Obsérvese que la primera correlación canónica R es 0.99, correspondiente a la correlación entre el primer par de variables canónicas, L(1). El valor R^2=0.98 indica que el 98% de la variabilidad de los datos es explicada por dicha correlación. La prueba para la hipótesis que la primera correlación canónica y todas las restantes son iguales a cero en la población, se basa en el estadístico lambda con 9 grados de libertad. En este ejemplo, el valor del estadístico (68.24) se asocia con un valor *p* menor a 0.001. Luego, la primera correlación canónica entre las calificaciones obtenidas en materias cuantitativas y no cuantitativas es significativamente distinta de cero en la población. La segunda correlación canónica, R=0.60, y las correlaciones menores no son significativamente distintas de cero como se puede visualizar a partir de los restantes valores *p*. En síntesis, una correlación canónica sería suficiente para medir la asociación, a nivel de calificaciones, entre ambos tipos de materias.

Nociones teóricas sobre análisis de correlaciones canónicas

El análisis de correlaciones canónicas (ACC) (Hotelling, 1936) se utiliza para determinar la relación lineal entre dos grupos de variables métricas unas consideradas como variables dependientes y otras como independientes. Vale decir que el análisis aborda el estudio de la *asociación entre dos conjuntos o grupos de variables*. Por ejemplo, supóngase que se tienen variables que indican la liquidez de firmas comerciales y otras variables que indican la contribución impositiva de esas mismas firmas, el ACC permite identificar y cuantificar la asociación entre liquidez y contribución impositiva, siendo estas dos características no medidas directamente sino a través de las variables que conforman cada grupo.

El ACC provee una medida de correlación entre una *combinación lineal* de las variables en un conjunto (en el ejemplo, una combinación lineal de las variables que miden liquidez), con una *combinación lineal* de las variables en el otro conjunto (combinación de variables que miden contribución impositiva).

En un primer paso del análisis, InfoStat determina el par de combinaciones lineales con máxima correlación. En un segundo paso, identifica el par con máxima correlación entre todos los pares no correlacionados con el par de combinaciones seleccionadas en el primer paso y así sucesivamente. Las combinaciones lineales de un par son llamadas *variables canónicas* y la correlación entre ellas, es llamada *correlación canónica*.

Para interpretar las variables canónicas, recordemos que el coeficiente de correlación simple entre dos variables, *Y* y *X*, (coeficiente producto momento de Pearson) se define como:

$$r_{12} = corr(Y, X) = \frac{Cov(Y, X)}{\sqrt{Var(Y)Var(X)}} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$$

Luego, si **x** es un vector de *q* variables y ***l***'**x** es una combinación lineal de **x**, la correlación entre una variable dependiente *Y* y la combinación ***l***'**x** será:

$$r_{y,l'x} = corr(Y, l'\mathbf{x}) = \frac{Cov(Y, l'\mathbf{x})}{\sqrt{Var(Y)Var(l'\mathbf{x})}}$$

El vector ***l*** que maximiza la correlación anterior es la combinación lineal resultante de ajustar un modelo de regresión múltiple de *Y* sobre **x** y se puede demostrar que:

$$r_{y,l'x} = corr(Y, l'\mathbf{x}) = \sqrt{R^2}$$

donde R^2 es el coeficiente de determinación de la regresión múltiple de **y** sobre las *q* variables en **x**. Luego el análisis de regresión múltiple es un caso particular del ACC donde uno de los conjuntos de variables a correlacionar tiene un solo elemento. El análisis de regresión simple correspondería a otro caso del ACC donde ambos conjuntos tiene sólo un elemento.

Ahora, si **y** es un vector de *p* variables, **x** es un vector de *q* variables y $\boldsymbol{l}_1$'**y** y $\boldsymbol{l}_2$'**x** son dos combinaciones lineales, la correlación entre dichas combinaciones es:

$$r_{l_1'y,l_2'x} = corr(l_1'\mathbf{y}, l_2'\mathbf{x}) = \frac{Cov(l_1'\mathbf{y}, l_2'\mathbf{x})}{\sqrt{Var(l_1'\mathbf{y})Var(l_2'\mathbf{x})}}$$

El ACC se basa en la obtención de los vectores $\boldsymbol{l}_1$ y $\boldsymbol{l}_2$ de modo que la correlación entre ambas combinaciones lineales de interés es máxima (correlación canónica).

Para encontrar los coeficientes de tales combinaciones lineales, se realiza la descomposición por valor singular de una matriz conformada por el producto de las matrices de varianzas y covarianzas de ambos conjuntos de variables.

Si denotamos por $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ a la matriz de varianzas y covarianzas del vector particionado $\begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix}$, las correlaciones canónicas (al cuadrado) ordenadas de mayor a menor son los autovalores (ordenados de mayor a menor) de la matriz $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ y los vectores de coeficientes de las combinaciones lineales relacionadas a $\mathbf{y}$, *i.e* vectores l_1, son obtenidos a partir de los autovectores, $\mathbf{e}_1$, de esa matriz, haciendo $\boldsymbol{l}_1' = e_1'\Sigma_{11}^{-1/2}$. Los vectores de coeficientes de combinaciones lineales $\boldsymbol{l}_2$ provienen de los autovectores de la descomposición de $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$ y los coeficientes canónicos $\boldsymbol{l}_2' = e_2'\Sigma_{22}^{-1/2}$. Los coeficientes de las combinaciones lineales se denominan pesos canónicos. Comúnmente se normalizan de manera tal que la variable canónica tenga varianza unitaria.

La primera correlación canónica nunca es menor que la mayor de las correlaciones múltiples entre cualquier variable y otra del grupo opuesto. Podría pasar que la primera correlación canónica sea muy alta mientras que todas las correlaciones múltiples para predecir una variable desde el conjunto opuesto sean pequeñas.

El ACC asume correlación del tipo lineal, otras correlaciones pueden pasar desapercibidas y distorsionar el análisis. La incorporación y eliminación de variables puede modificar sustancialmente el análisis al igual que la presencia de puntos influyentes. Técnicas de diagnóstico comunes en el análisis de regresión pueden ser utilizadas para la identificación de puntos influyentes. En el ACC no se requiere normalidad multivariada a menos que se pretendan obtener errores estándares y pruebas de hipótesis para las correlaciones.

El número de correlaciones canónicas que puede ser extraído desde estas descomposiciones es igual al mínimo de los números p y q (cardinalidad de cada uno de los conjuntos de variables que se desean correlacionar). Los coeficientes de correlación canónica al cuadrado representan la proporción de la varianza total explicada por cada variable canónica. Usualmente se reporta bajo el nombre de "estructura canónica total" a las correlaciones simples entre las variables respuestas y las variables canónicas.

Regresión por Mínimos Cuadrados Parciales

Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ PLS (del inglés, *Partial Least Squares*) es un método estadístico multivariado relativamente nuevo. Es una técnica que generaliza y combina el ACP y el análisis de Regresión Lineal. Es particularmente útil cuando se desea predecir un conjunto de variables dependientes (Y) desde un conjunto (relativamente grande y posiblemente correlacionadas) de variables predictoras (X). El objetivo del método PLS es describir Y a partir de X y su estructura de variación común.

Cuando hay más observaciones que variables predictoras y no existe problema de multicolinealidad, la predicción de Y en función de X puede realizarse eficientemente con un análisis de regresión lineal múltiple. PLS se usa cuando existe correlación entre las variables predictoras y/o existen más predictoras que observaciones. El problema de la estimación en estos casos podría resolverse combinando linealmente las predictoras con un ACP y luego regresionando Y con un número reducido de componentes principales. Pero hay que recordar que las CP explican variación en X y nada nos dicen sobre la relación de Y con X. Por el contrario la técnica PLS busca una solución óptima o de compromiso entre el objetivo de explicar la máxima variación en X y encontrar las correlaciones de éstas con Y.

Si llamamos X e Y a los dos conjuntos de variables y suponemos que el número de variables en X es m (X_1, X_2, ...,X_m) y el número de variables en Y es n (Y_1, Y_2, ...,Y_n), es posible construir una matriz R de correlación tal que su elemento R_{ij} sea la correlación entre X_i (i=1,..m) e Y_j (j=1,..,n). Esta matriz no tiene unos en la diagonal y usualmente no es cuadrada. La idea en PLS es obtener un vector de m coeficientes A_i, uno para cada variable en X y un vector de n coeficientes B_j, uno para cada variable en Y, tal que el producto AB' (*i.e.*, matriz cuya entrada ij es A_i*B_j) aproxime bien a la matriz R en el sentido mínimo cuadrático (i.e., minimizando la suma de los términos $(R_{ij}-A_i*B_j)^2$). Podríamos decir que estos coeficientes permiten combinar las variables de cada conjunto para explicar la variabilidad debida a la relación o correlación entre ambos bloques.

Una aplicación clásica de PLS es extender la regresión múltiple cuando existe correlación entre las predictoras o como se indicó anteriormente cuando el número de observaciones es pequeño en relación al número de regresoras. La implementación de PLS en InfoStat esta orientada a la obtención de espacios de representación similares a los obtenidos con PCA pero involucrando un con junto adicional de covariables que se utilizan para explicar las relaciones que se visualizan en la representación bi-plot de un PCA entre los objetos y las variables. Los resultados de PLS, son presentados a través de un "tri-plot". Nos referimos a tri-plot cuando se dispone de un gráfico bi-plot sobre el que además se grafican covariables para explicar la asociación entre los "punto del espacio fila" y los "puntos del espacio columna" de un biplot.

Objetivos

Descubrir y reportar la naturaleza de las relaciones de variables predictoras con una o varias variables respuesta (*i.e.*, un conjunto de variables respuestas).

Datos

Se necesitan *I* observaciones o casos descriptos por *n* variables dependientes (bloque de variables Y) y *m* predictores colectados sobre estos *I* casos en una matriz de datos *I*×*m* (bloque de variables X). La tabla de datos en InfoStat deberá contener *I* casos y al menos (*m*+*n*) columnas.

Ejemplo 37: El archivo Factores limitantes soja.idb, se usa para ejemplificar una aplicación de PLS que explique la interacción genotipo-ambiente (GE) (campaña agrícola 01_02) en función de las siguientes covariables ambientales: Ra3, %MD, %pi, PrB2t y MO. En esta campaña intervinieron 3 genotipos (A5520RG, A6040RG y DM4800RR) y 7 localidades (ambientes) Cavanagh, Totoras, Oliveros, Maizales, Bouquet, Rueda, y C.Gómez.

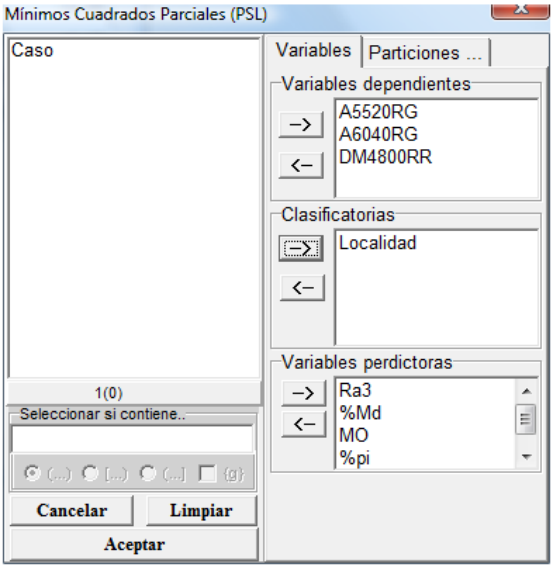
La matriz Y de este ejemplo contiene términos de interacción entre 7 localidades y 3 genotipos y la matriz X contiene las covariables ambientales antes descriptas. En la tabla se muestra en sombreado el bloque de la matriz X y no sombreado el bloque de la matriz Y (datos: Factores limitantes soja.idb).

EjemploPLS

Caso	Localidad	Ra3	%Md	MO	%pi	PrB2t	A5520RG	A6040RG	DM4800RR
1	Bouquet	493.03	17.00	2.87	52.50	28.00	13.10	-23.86	10.76
2	C.Gómez	488.83	0.00	3.13	21.67	14.67	-2.07	4.06	-1.99
3	Cavanagh	548.13	22.00	3.65	37.07	20.00	8.56	-10.28	1.71
4	Maizales	469.80	31.00	3.07	88.50	25.00	4.93	7.92	-12.85
5	Oliveros	452.43	19.00	2.54	59.07	31.33	-21.68	2.31	19.37
6	Rueda	368.03	3.57	2.85	35.33	28.67	-0.81	17.51	-16.70
7	Totoras	540.47	0.00	3.55	19.00	33.33	3.99	-2.99	-1.00

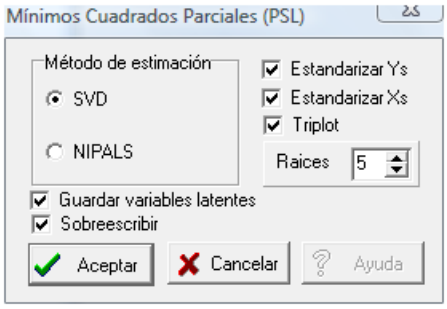
Categórica | registros: 8*9

En el Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ PLS, ventana **Mínimos cuadrados parciales (PSL)**, se seleccionaron en **Variables dependientes** los genotipos A5520RG, A6040RG y DM4800RR Como **Clasificatorias** la localidad y como **Variables predictoras** Ra3, %MD, MO, %pi y PrB2t. Estos comandos permiten **¡Error! No se encuentra el origen de la referencia.**la implementación de la rutina SVD para PLS y la obtención del tri-plot, ejecutados sobre esta tabla. Análisis PLS (las columnas de Y deben ir como variables dependientes, mientras que las filas como clasificatorias;



las columnas de X como predictoras).

Luego de activar el botón **Aceptar**, aparece la ventana **Mínimos Cuadrados Parciales (PLS)**, en la cual las opciones **SVD**, **Estandarizar Ys**, **Estandarizar Xs**, **Guardar variables latentes** y **Sobreescribir** están tildadas por defecto. Para este ejemplo se tildó también **Triplot** y **Raíces 5**, como puede verse en la ventana adjunta.



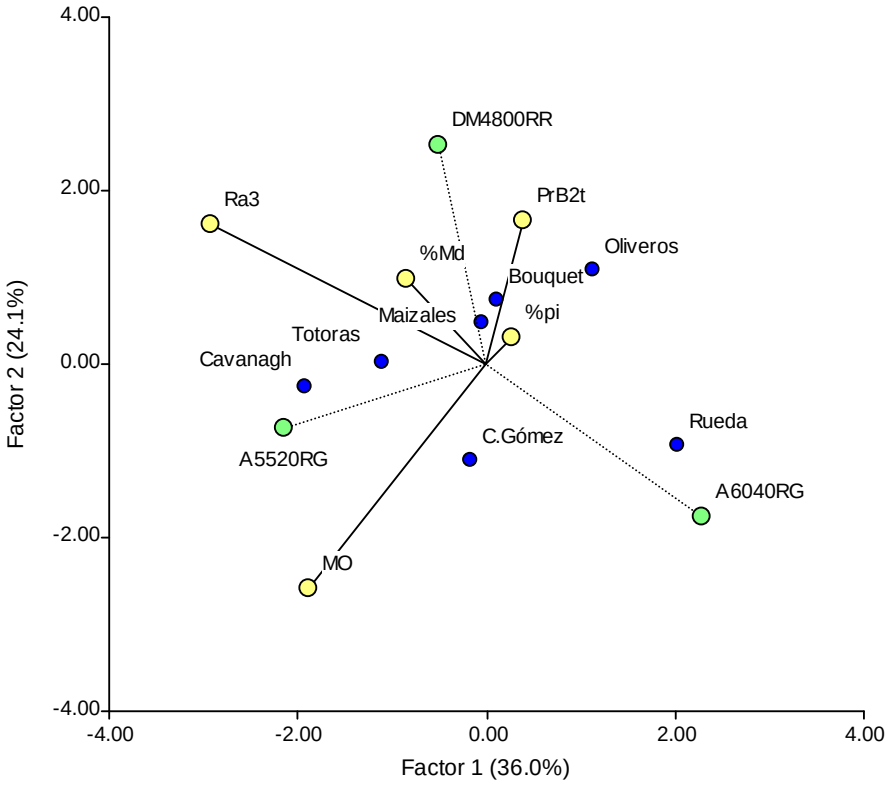


Figura 30: Tri-plot de la correlación entre una matriz de interacción entre 3 genotipos y 7 ambientes versus una matriz de 5 covariables ambientales. Archivo: Factores limitantes soja.idb.

La interacción GE se explica en su totalidad a partir de las dos primeras CP, según lo muestran los autovalores. Los *scores* de genotipos y ambientes para el estudio de interacción se presentan en la salida, ellos sirven para asociar genotipos con ambientes, pero no para explicar esta asociación con variables en X, las nuevas variables latentes obtenidas a partir de la técnica PLS se muestran en la ventana resultados (no presentada aquí). Al correlacionar la matriz de residuos del modelo AMMI(2) con las covariables ambientales, las covariables de mayor "inercia" sobre el eje 1 del tri-plot resultaron ser Ra3 y MO (los radios en la Figura presentada, correspondientes a estas variables tienen una gran proyección sobre el eje 1). Luego las interacciones detectadas en este conjunto de datos, desde el punto de vista ambiental, son principalmente atribuidas a estas dos variables.

Valores de Ra3, relativamente altos se registraron en Cavanagh y en Totoras, éstos podrían explicar el desempeño mejor que tuvo el genotipo A5520RG respecto a los otros en esas localidades. La MO también fue relativamente alta en Cavanagh y Totoras y muy baja en Oliveros. Las características de suelo no relacionadas con la MO, no resultaron importantes para explicar las interacciones en esta campaña. El cultivar A6040RG se desempeñó, relativo a los otros dos cultivares, mejor en Rueda y en Oliveros; la interacción con Rueda se correlaciona negativamente con Ra3. La segunda dimensión del tri-plot se asocia con las adaptaciones mejores de DM4800 en Oliveros que presenta un menor contenido de MO que los otros sitios.

Análisis de la varianza multivariado

Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DE LA VARIANZA MULTIVARIADO, permite probar hipótesis sobre igualdad de vectores de medias en dos o más poblaciones.

Cuando se estudian *p* variables para cada nivel de uno o más factores de diseño, el análisis de la varianza multivariado (MANOVA), se utiliza para realizar inferencias simultáneas sobre los efectos de los factores del modelo de análisis. Los modelos de análisis pueden involucrar tanto factores de clasificación como covariables (variables continuas). Los factores de clasificación pueden estar cruzados o anidados y la expresión del lado derecho de la ecuación del modelo se escribe en InfoStat siguiendo las mismas pautas establecidas para el análisis univariado de varianza (ver Análisis de la varianza). A diferencia del análisis de varianza univariado, en este módulo se deberán seleccionar más de una variable dependiente. Para análisis de varianza involucrando varias variables, un valor faltante en una de las variables dependientes elimina la observación completa.

InfoStat provee automáticamente cuatro estadísticos diferentes para pruebas de hipótesis multivariadas. Para cada estadístico se reportan también las aproximaciones F (Johnson y Wichern, 1998). InfoStat permite definir matrices específicas para probar hipótesis relacionadas a diferencias entre grupos (niveles de un factor de clasificación) para cada una de las variables dependientes tanto como para combinaciones lineales de dichas variables. La hipótesis lineal general multivariada se expresa como: H_0: **CBA=0**, donde **C** permite formular contrastes entre las filas de **B** (por ejemplo el efecto grupo) y **A** permite definir nuevas variables respuesta a partir de combinaciones lineales de las columnas de **B**. Si no se

especifica una matriz **A** particular, InfoStat asume que **A** es la matriz identidad, caso contrario las pruebas se realizan para las variables transformadas definidas por las columnas de la matriz **A**. En la ventana Análisis de la varianza multivariado se declararán las Variables dependientes y la Variable de clasificación. Al Aceptar aparece una ventana con las siguientes solapas: Modelo, Comparaciones, Contrastes/comb. lineales y Combinaciones lineales columnas.

En la solapa **Modelo**, en **Especificación de los términos del modelo** se puede escribir el modelo deseado, en **Variables de clasificación** aparecen las variables del lado derecho de la ecuación del modelo declaradas en la ventana previa. En **Covariables** se elegirán las variables que se deseen incluir como tales.

En la solapa **Comparaciones**, se puede seleccionar un método de comparación *a posteriori*. En **Medias a comparar** se elige el factor para el cual se compararán las medias o bien **Todas las medias** (en caso de diseños a una sola vía de clasificación, las dos opciones producirán idénticos resultados). Se puede decidir, al igual que en el módulo univariado de ANAVA, la **Presentación** en forma de lista o matricial de las medias a comparar y el **Nivel de significación** de la prueba. Se dispone de las pruebas de Hotelling y Hotelling corregida por la desigualdad de Bonferroni para realizar comparaciones de vectores medios entre grupos. Si el usuario no declara una matriz de transformaciones de las variables respuestas, el perfil de cada grupo estará conformado por todas las variables dependientes seleccionadas para el análisis. En caso de existir una matriz **A** distinta a la identidad, los vectores a comparar se conforman a partir de las variables transformadas.

En la solapa **Contrastes/comb. lineales** se especifica la matriz **C**, el usuario debe seleccionar el factor cuyos niveles desean ser contrastados en **Seleccionar términos**. Al realizar esta acción InfoStat permitirá visualizar los niveles del factor seleccionado en la ventana **Tratamientos**. En **Filas de C de la hipótesis H_0: CBA=0** se deberá ingresar la matriz **C**, es decir los coeficientes de los contrastes o combinaciones lineales sobre el factor seleccionado que se requieren. Cuando se ingresan contrastes InfoStat permite **controlar la ortogonalidad** de los mismos si el usuario lo requiere.

En la solapa **Combinaciones lineales entre columnas**, se especifica la transpuesta de la matriz **A** de la hipótesis lineal general multivariada. En **Designación de las columnas de B**, el usuario puede visualizar las variables dependientes en estudio para facilitar la especificación de **A**.

Para los datos del archivo *Iris* se presenta a continuación la salida del análisis multivariado de varianza para probar la hipótesis de igualdad de vectores medios entre las tres especies ("tratamientos"). Se realizó también un contraste para comparar la especie 2 con la 3. Se invocó el menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DE LA VARIANZA MULTIVARIADO, en la ventana **Análisis de la varianza multivariado** se designaron las variables respuesta "SepalLen", "SepalWid", "PetalLen" y "PetalWid" como **Variables dependientes** y la variable Especie fue seleccionada como **Variable de clasificación**. En la solapa **Contrastes/c.lineales**, **Seleccionar factor**, se eligió "Especie". En **Tratamientos** aparece 1,2,3 (los tres niveles del factor seleccionado). En **Filas de C de la hipótesis H_0: CBA=0** se escribió 0 1 –1 para contrastar el vector de medias de la especie 2 con el de la especie 3.

Tabla 55: Análisis de la varianza multivariado.

Cuadro de análisis de la varianza (Wilks)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Especie	0.02	199.15	8	288	<0.0001

Cuadro de análisis de la varianza para contrastes(Wilks)

Especie	Estadístico	F	gl(num)	gl(den)	p
Contraste1	0.25	105.31	4	144	<0.0001
Total	0.25	105.31	4	144	<0.0001

Cuadro de análisis de la varianza (Pillai)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Especie	1.19	53.47	8	290	<0.0001

Cuadro de análisis de la varianza para contrastes(Pillai)

Especie	Estadístico	F	gl(num)	gl(den)	p
Contraste1	0.75	105.31	4	144	<0.0001
Total	0.75	105.31	4	144	<0.0001

Cuadro de análisis de la varianza (Lawley-Hotelling)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Especie	32.48	580.53	8	286	<0.0001

Cuadro de análisis de la varianza para contrastes(Lawley-Hotelling)

Especie	Estadístico	F	gl(num)	gl(den)	p
Contraste1	2.93	105.31	4	144	<0.0001
Total	2.93	105.31	4	144	<0.0001

Cuadro de análisis de la varianza (Roy)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Especie	32.19	1166.96	4	145	<0.0001

Cuadro de análisis de la varianza para contrastes(Roy)

Especie	Estadístico	F	gl(num)	gl(den)	p
Contraste1	2.93	105.31	4	144	<0.0001
Total	2.93	105.31	4	144	<0.0001

Coeficientes de los contrastes

Especie	C'(1)
1	0.00
2	1.00
3	-1.00

Las dos primeras tablas corresponden al análisis de varianza multivariado para probar la igualdad de vectores medios entre las tres especies y el contraste entre las especies 2 y 3 en base al estadístico de Wilks (0.02 y 0.25, respectivamente). También se observan los valores de la aproximación F para ambas pruebas (199.15 y 105.31), los grados de libertad para el estadístico F y los niveles de probabilidad asociados a cada prueba. Como se puede observar en ambos casos el valor *p* es menor a 0.001, es decir existen diferencias estadísticamente significativas entre los centroides de las observaciones multivariadas de las tres especies y entre los centroides de las observaciones de la especie 2 con respecto al de la especie 3. Las pruebas realizadas involucran de forma conjunta parámetros de las 4 variables dependientes, permitiendo así probar la igualdad de grupos a partir de las 4 variables simultáneamente. De la misma manera deben analizarse los tres pares de tablas restantes, que presentan las mismas pruebas de hipótesis pero realizadas en base a los estadísticos de Pillai, Lawley-Hotelling y Roy. Finalmente InfoStat reporta la matriz **C** especificada por el usuario, para la realización de uno o más contrastes (en este ejemplo **C** contiene sólo un contraste).

Nociones teóricas sobre análisis multivariado de varianza

El modelo lineal subyacente puede expresarse en términos matriciales como sigue:

$$\mathbf{Y} = \mathbf{XB} + \varepsilon$$

donde **Y** es una matriz $n \times p$ con n=número de observaciones o casos del archivo y p=número de variables seleccionadas como dependientes, **X** es una matriz $n \times k$ con k=número de parámetros fijos del modelo asociados a una variable, **B** una matriz $k \times p$ conteniendo los parámetros fijos asociados a las p variables y ε una matriz $n \times p$ de términos de error. Se supone que las n filas de ε son independientes y tienen distribución normal p-variada con matriz de varianzas y covarianzas Σ de dimensión $p \times p$ para cada caso. Luego colectando todos los términos de error de la matriz ε en un vector, *vec*(ε), InfoStat asume que: $vec(\varepsilon) \sim N(\mathbf{0}, \mathbf{I}_n \otimes \Sigma)$, donde *vec*(ε) es la función que arregla en forma vectorial los elementos de la matriz **ε** y el símbolo ⊗ denota producto Kronecker; Σ es la matriz de varianzas y covarianzas entre las variables respuesta dentro de cada observación. Luego, en el MANOVA se realiza el supuesto de que los términos de error son independientes entre observaciones pero no entre variables. La normalidad multivariada es necesaria para probar hipótesis. La matriz Σ es estimada por $\mathbf{S} = \left[(\mathbf{e'e})/(n-r)\right] = \left[\left(\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)\right)/(n-r)\right]$ donde $\hat{\mathbf{B}} = (\mathbf{X'X})^{-}\mathbf{X'Y}$, **r** es el rango de la matriz **X** y **e** es la matriz de residuos.

Si **S** es escalada para lograr que los elementos de la diagonal sean unos, los restantes elementos de **S** se denominan correlaciones parciales de las variables dependientes ajustando por las variables del lado derecho del modelo. InfoStat permite obtener esta matriz a solicitud del usuario.

Matrices específicas de contrastes y de combinaciones lineales de las variables respuesta y de los efectos del modelo permiten probar una amplia gama de hipótesis. La hipótesis lineal general multivariada se expresa como: H_0: **CBA=0**, donde **C** permite formular contrastes entre las filas de **B** (por ejemplo el efecto grupo) y **A** permite definir nuevas variables respuesta a partir de combinaciones lineales de las columnas de **B**.

Las pruebas estadísticas multivariadas se realizan a partir de la estimación de las matrices $\mathcal{H}$ y $\mathcal{E}$ correspondientes a las sumas de cuadrados y productos cruzados asociadas a la hipótesis y al error, respectivamente. Cuando un factor del modelo tiene g niveles, la estimación de la matriz de la hipótesis sobre igualdad de vectores medios entre los grupos de observaciones definidas por los niveles de dicho factor es $\mathbf{H} = \sum_{i=1}^{g} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$ donde $\bar{\mathbf{x}} = \frac{1}{g}\sum_{i=1}^{g} \bar{\mathbf{x}}_i$ y la matriz residual de sumas de cuadrados y productos cruzados de los términos de error es $\mathbf{E} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' = \sum_{i=1}^{g} (n_i - 1)\mathbf{S}_i$. Estas matrices cumplen la misma misión que el numerador y el denominador del estadístico F univariado, *i.e.* la matriz **H** provee estimaciones de la variación (y covariación) entre grupos y la matriz **E** provee estimaciones de la variación (y covariación) dentro de los grupos. Las matrices **H** y **E** son construidas a partir de las siguientes expresiones generales:

$\mathbf{H} = \mathbf{A}'(\mathbf{C}\hat{\mathbf{B}})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{B}})\mathbf{A}$ $\mathbf{E} = \mathbf{A}'(\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'(\mathbf{X}'\mathbf{X})^{-}\hat{\mathbf{B}})\mathbf{A}$

InfoStat provee cuatro estadísticos para cada prueba multivariada, todos ellos funciones de los autovalores de $\mathbf{E}^{-1}\mathbf{H}$ (o $(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}$) (Pillai, 1960). Estos son:

Lambda de Wilks=det($\mathbf{E}$)/det($\mathbf{H}$+$\mathbf{E}$)

Traza de Pillai=traza($\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$)

Traza de Hotelling-Lawley=traza($\mathbf{E}^{-1}\mathbf{H}$)

Máximo autovalor de Roy=λ, es el mayor autovalor de $\mathbf{E}^{-1}\mathbf{H}$

Medidas repetidas en el tiempo (enfoque multivariado)

Al medir una variable sobre la misma unidad experimental en distintos momentos de tiempo se obtienen mediciones seriales que se caracterizan por estar correlacionadas ya que las mismas acarrean un mismo efecto de unidad experimental. El procedimiento de recolección de información sobre la misma unidad produce una serie de medidas repetidas sobre cada unidad. Pueden existir o no otros factores que se reconozcan como fuentes de variabilidad entre las unidades o sujetos experimentales (*between subject factor*), pero siempre en este tipo de estudios se reconoce el factor tiempo como aquel con potencialidades de introducir variabilidad entre las observaciones registradas dentro de la misma unidad (*within subject factor*).

Las medidas repetidas en el tiempo pueden analizarse como perfiles multivariados, donde las respuestas observadas en cada momento de tiempo representan las variables de análisis. Es decir cada observación se corresponde con un vector t-dimensional donde *t* representa el número de instancias en el tiempo en que se registra el valor de la variable de análisis. El análisis multivariado de medidas repetidas permite modelar las correlaciones existentes entre observaciones seriales.

Las hipótesis clásicas a probar en un análisis de mediciones repetidas en el tiempo con un factor tratamiento agrupando las observaciones son: 1) no hay interacción tiempo×tratamiento, 2) no hay efecto tiempo y 3) no hay efecto tratamiento o grupo. A través del uso de matrices **A** y **C** específicas se pueden probar dichas hipótesis en InfoStat.

Usando el módulo análisis de varianza (univariado), se podría ajustar un modelo de parcelas divididas a los datos de experimentos con mediciones repetidas. En dicho modelo el factor tratamiento se asocia a las unidades experimentales de mayor tamaño y el factor tiempo a las "subparcelas" (Winer, 1971; Morrison, 1976). Los sujetos o unidades anidadas dentro del factor principal constituyen el término de error para el factor tratamiento. El análisis es apropiado sólo si las medidas no están correlacionadas en el tiempo (modelo esférico) o se puede sostener el supuesto de igual correlación entre cualquier par de mediciones repetidas sobre un mismo individuo (modelo de simetría compuesta) para las matrices de varianzas y covarianzas de las observaciones dentro de una misma unidad experimental. Si las matrices de covarianza dentro del sujeto tienen dichas características se dice que cumplen con la condición Huynh-Feldt (Huynh-Feldt, 1970). Por otro lado, la aproximación multivariada para el análisis de mediciones repetidas (Cole y Grizzle, 1966) no asume ningún modelo

particular para dicha matriz, sino que se basa en la estimación de todas las covarianzas posibles entre las mediciones repetidas. Este modelo desestructurado, poco parsimonioso, debe ser abordado cuando se tienen suficientes observaciones para la estimación de parámetros. En general se pide que el número de observaciones repetidas sea menor o igual al número de repeticiones del experimento. Estructuras de covarianza intermedias entre la simetría compuesta y la desestructurada pueden proveer soluciones de interés en la práctica de este tipo de mediciones (aproximación basada en modelos mixtos). En la versión 2008 de InfoStat, existe un procedimiento para ajustar modelos mixtos. Este procedimiento usa al lenguaje R como motor de cálculos y construye una interfase con las rutinas lme y gls de la librería nlme. Un documento explicativo de este modulo se puede encontrar en el siguiente link. http://www.infostat.com.ar/descargas/demo/MMixtosInfoStat.pdf

La organización de la tabla de datos para realizar el análisis de mediciones repetidas, a través de la aproximación multivariada, debe corresponderse a la estructura solicitada por el MANOVA. Los registros de la variable realizados en diferentes momentos de tiempo deben aparecer en diferentes columnas. Luego, además de las columnas asociadas a criterios de clasificación y/o covariables, la tabla tendrá al menos un número *t* adicional de columnas.

Ejemplo 38: se tienen dos tratamientos donde se registra una variable V en 3 momentos de tiempo sobre tres sujetos por tratamiento, el archivo deberá tener el siguiente formato:

Tabla 56: Organización de datos para análisis de mediciones repetidas. Archivo MedRep.

Trat	Suj	V1	V2	V3
1	1	125	182	206
1	2	107	178	201
1	3	167	182	223
2	1	167	199	250
2	2	163	208	217
2	3	173	192	233

En InfoStat, las hipótesis de interés pueden ser probadas a través de dos corridas del análisis multivariado de varianza:

Corrida 1: Para los datos del archivo *MedRep* (datos de la tabla anterior) se invocó el menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DE LA VARIANZA MULTIVARIADO. En la ventana **Análisis de la varianza multivariado** se designaron las variables respuesta V1, V2 y V3 como **Variables dependientes** y la variable "Trat" fue seleccionada como **Variable de clasificación**. En la solapa **Contrastes/comb. lineales**, **Seleccionar factor**, se eligió "Trat". En **Tratamientos** aparece 1 y 2 (los dos niveles del factor seleccionado). En **Filas de C de la hipótesis H_0: CBA=0** se escribió "1 1" y se activó el casillero **Comb.lineales** para indicar que se acaba de ingresar una combinación lineal en lugar de un contraste. El número de unos a ingresar deberá ser igual al número de niveles del factor seleccionado, la ventana tratamientos facilita recordar dicho número. Finalmente, se activa la solapa **Combinaciones lineales columnas** y en el espacio destinado a escribir la matriz **A'** se ingresan las siguientes dos filas de contrastes: 1 –1 0 y 0 1 –1. Las listas de variables activas aparecerán bajo el título **Designación de las columnas de B**. La matriz **A'** deberá tener tantas columnas como variables se muestren en dicha lista, el número de filas es *t-1*. Se obtendrán los siguientes resultados (se ha dejado sólo los resultados

correspondientes al estadístico de Wilks, la interpretación es idéntica para los otros tres estadísticos reportados por InfoStat):

Tabla 57: Análisis de medidas repetidas (corrida 1). Archivo MedRep.

Cuadro de análisis de la varianza (Wilks)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Trat	0.83	0.31	2	3	0.7560

Cuadro de análisis de la varianza para combinaciones lineales(Wilks)

Trat	Estadístico	F	gl(num)	gl(den)	p
Fila 1 de C	0.04	41.30	2	3	0.0066
Total	0.04	41.30	2	3	0.0066

Cuadro de análisis de la varianza (Pillai)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Trat	0.17	0.31	2	3	0.7560

Cuadro de análisis de la varianza para combinaciones lineales(Pillai)

Trat	Estadístico	F	gl(num)	gl(den)	p
Fila 1 de C	0.96	41.30	2	3	0.0066
Total	0.96	41.30	2	3	0.0066

Cuadro de análisis de la varianza (Lawley-Hotelling)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Trat	0.20	0.31	2	3	0.7560

Cuadro de análisis de la varianza para combinaciones lineales(Lawley-Hotelling)

Trat	Estadístico	F	gl(num)	gl(den)	p
Fila 1 de C	27.53	41.30	2	3	0.0066
Total	27.53	41.30	2	3	0.0066

Cuadro de análisis de la varianza (Roy)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Trat	0.20	0.31	2	3	0.7560

Cuadro de análisis de la varianza para combinaciones lineales(Roy)

Trat	Estadístico	F	gl(num)	gl(den)	p
Fila 1 de C	27.53	41.30	2	3	0.0066
Total	27.53	41.30	2	3	0.0066

Coeficientes de la matriz C'

Trat	C'(1)
1	1.00
2	1.00

Matriz A

Define combinaciones lineales de las columnas de la matriz de parámetros B

Variable	Col(1)	Col(2)
V1	1.00	0.00
V2	-1.00	1.00
V3	0.00	-1.00

La hipótesis de interacción tiempo y tratamiento es contrastada a partir de la primera tabla donde la fuente de variación viene dada por el factor tratamiento. Debido a la combinación lineal propuesta de las variables dependientes, a través de la matriz **A**, el estadístico de Wilks, λ=0.83, que puede aproximarse al valor 0.31 de una F con 2 y 3 grados de libertad, permite probar la hipótesis de interacción tiempo-tratamiento. El valor *p*=0.7560 sugiere que no existen evidencias para rechazar la hipótesis de falta de interacción, es decir las diferencias entre tratamientos no cambian con el tiempo. La segunda tabla se asocia a la

hipótesis de falta de efecto tiempo, esta hipótesis se prueba a partir de las diferencias en la variable respuesta entre dos momentos de tiempos subsecuentes (integrada sobre todos los tratamientos). La probabilidad de obtener valores del estadístico de Wilks mayores al observado es p=0.0066, por lo tanto se rechaza la hipótesis de falta de efecto tiempo. Luego los perfiles de ambos tratamiento cambian con el tiempo. Aún no se sabe cuál es la tendencia que existe en el tiempo, sólo se conoce que la respuesta no es constante. Al final de la salida InfoStat provee las matrices **C** y **A** especificadas por el usuario.

Corrida 2: Tiene por objetivo probar la hipótesis de igualdad de medias de tratamiento. La misma se prueba integrando todas las respuestas de un tratamiento en el tiempo a través de la matriz **A**. Para ello volver a la ventana **Análisis de la varianza multivariado** dejando la misma selección que la corrida anterior. Limpiar la solapa **Contrastes/comb. lineales** (en **Filas de C de la hipótesis H_0: CBA=0** no debe quedar nada). En la solapa **Combinaciones lineales columnas** y en el espacio destinado a escribir la matriz **A'** se ingresan una sola fila de unos (tantos unos como variables aparecen bajo el título **Designación de las columnas de B**, notar que **A** es el vector unitario). Se obtendrán los resultados de la Tabla 58 (se muestran sólo los resultados correspondientes al estadístico de Wilks, la interpretación es idéntica para los otros tres estadísticos reportados por InfoStat). Los resultados son idénticos para los cuatro estadísticos multivariados ya que la combinación de las columnas en una única variable establece una prueba univariada y por tanto el estadístico F es exacto. En el presente ejemplo, las diferencias entre tratamientos son significativas al 5%.

Tabla 58: Análisis de medidas repetidas (corrida 2). Archivo MedRep.

Cuadro de análisis de la varianza (Wilks)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Trat	0.32	8.33	1	4	0.0447

Cuadro de análisis de la varianza (Pillai)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Trat	0.68	8.33	1	4	0.0447

Cuadro de análisis de la varianza (Lawley-Hotelling)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Trat	2.08	8.33	1	4	0.0447

Cuadro de análisis de la varianza (Roy)

F.V.	Estadístico	F	gl(num)	gl(den)	p
Trat	2.08	8.33	1	4	0.0447

Matriz A

Define combinaciones lineales de las columnas de la matriz de parámetros B

Variable	Col(1)
V1	1.00
V2	1.00
V3	1.00

Para pruebas que involucran sólo un factor principal (*between subject factor*) la aproximación univariada de parcelas divididas y la multivariada producen los mismos resultados. Las diferencias de niveles de probabilidad se producen en pruebas que involucran el factor tiempo (*within subject factor*). La matriz **A** utilizada para la prueba de hipótesis del efecto tiempo en el ejemplo anterior, produce una transformación de las variables dependientes en diferencias entre los registros de la variable en tiempos sucesivos. Esta transformación produce el análisis usualmente conocido como **análisis de perfiles**

(Johnson y Wichern, 1998). Otras matrices, las cuales no cambian los resultados generales de las pruebas mencionadas, pueden ser propuestas para analizar la naturaleza de las respuestas en el tiempo. Por ejemplo, si un nivel del factor tiempo puede verse como un control o una lectura de referencia se podría utilizar una matriz **A** que permitiera estudiar el efecto tiempo a través del **contraste de todos los tiempos con el tiempo de referencia**.

$$A' = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

Examinando cada uno de los contrastes implementados el usuario puede identificar los momentos del tiempo en que se produjeron respuestas diferentes al control.

Con *t* niveles de tiempo, es posible probar *t-1* tendencias lineales a partir de una matriz **A** que contenga los coeficientes de los **polinomios ortogonales** para tal fin. Para el ejemplo anterior, la siguiente matriz permite probar si los cambios en el tiempo responden a una tendencia lineal y/o cuadrática.

$$A' = \begin{bmatrix} -1 & 0 & 1 \\ -0.5 & 1 & -0.5 \end{bmatrix}$$

La matriz de transformación de variables respuestas conocida como matriz Helmert compara cada nivel (tiempo) de la variable respuesta con la media de los niveles subsecuentes. Este tipo de matriz es conveniente para situaciones experimentales donde interesa identificar el momento de tiempo en el cual la respuesta se estabiliza (deja de cambiar). En el ejemplo en cuestión esta transformación se debería representar por la siguiente matriz:

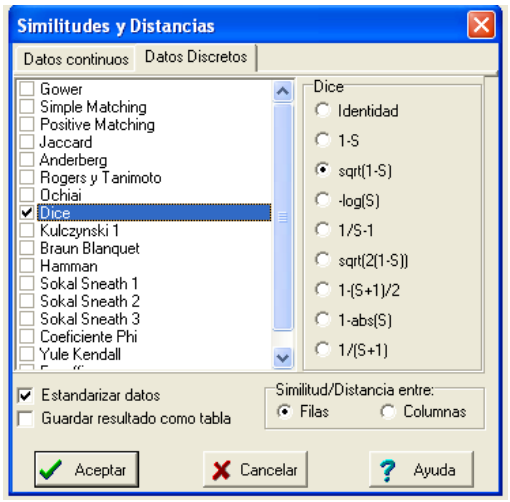
$$A' = \begin{bmatrix} 1 & -0.5 & -0.5 \\ 0 & 1 & -1 \end{bmatrix}$$

Para determinar el punto (momento de tiempo) en el cual se produce un "pico" (positivo o negativo) en el perfil de las observaciones en el tiempo, se debería realizar una análisis de varianza para las variables transformadas que representan la diferencia en la respuesta entre momentos de tiempos adyacentes.

Correlación-distancias-similitudes

Menú ⇒ ANÁLISIS MULTIVARIADO ⇒ CORRELACION-DISTANCIAS-SIMILITUDES permite obtener para un conjunto de variables seleccionadas medidas de correlación-asociación y distancias. En el **Selector de variables** se pueden especificar las variables de interés y uno o más **Criterios de clasificación**. Si no tiene variables de clasificación seleccione "caso". Al **Aceptar** aparecerá una ventana de diálogo en la que se puede elegir entre dos solapas: **Datos continuos y Datos discretos**. En las dos solapas se pueden **estandarizar los datos** y elegir un análisis por **filas** o por **columnas.**

Se pueden elegir los estadísticos para generar matrices de **distancia**. InfoStat produce una matriz $n \times n$ o $p \times p$ según se indique análisis por filas o por columnas, respectivamente. Los elementos de la matriz reportada son medidas de distancia entre puntos filas o columnas obtenidas a partir del estadístico seleccionado. Las distancias ofrecidas son: Euclídea; Euclídea promedio; Euclídea cuadrado; Manhattan; Manhattan promedio; Bray-Curtis; Bray-Curtis promedio (Camberra) y Excoffier. Además, se pueden calcular distancias como funciones de medidas de similitud. Las medidas de **similitud** ofrecidas para la obtención de distancias son: Gower, correlación de Pearson, correlación de Spearman, simple matching, positive matching, Jaccard, Anderberg, Rogers y Tanimoto, Ochiai, Dice, Kulczynsky 1, Braun Blanquet, Hamman, Sokal Sneath 1, Sokal Sneath 2, Sokal Sneath 3, Coeficiente Phi, Yule Kendall. Cuando se seleccione una de estas similitudes, automáticamente aparecerá una subventana a la derecha para seleccionar la función a usar para transformar las similitudes en distancias. Todas son funciones de S, donde S es la similitud seleccionada.



Nociones teóricas sobre medidas de distancia y asociación

Una medida de distancia métrica entre dos puntos, digamos P y Q, satisface los siguientes requerimientos:

$d(P,Q) = d(Q,P)$ la distancia es simétrica,

$d(P,Q) > 0$ si $P \neq Q$,

$d(P,Q) = 0$ si $P = Q$, y

$d(P,Q) \leq d(P,R) + d(R,Q)$, desigualdad triangular.

Las medidas de distancia pueden ser convertidas a medidas de similitud entre observaciones. En las medidas de similitud, al contrario de las de distancia, el valor que se obtiene es tanto mayor cuanto más próximos están los elementos considerados. Así como las medidas de distancia poseen ciertas propiedades, las medidas de similitud deben cumplir:

$$0 \leq s(P,Q) \leq 1$$

$$s(P,P) = 1$$

$$s(P,Q) = s(Q,P)$$

La distancia **Euclídea** o distancia en línea recta desde una observación p-dimensional, al origen es,

$$d(O,x) = L(x) = \sqrt{x_1^2 + x_2^2 + ...x_p^2}$$

y representa una generalización del Teorema de Pitágoras que establece la distancia entre un punto y el origen en el plano,

$$d(O,x) = \sqrt{x_1^2 + x_2^2}$$

La distancia Euclídea entre dos puntos arbitrarios del espacio p-dimensional es la raíz cuadrada de la suma de *p* diferencias (al cuadrado) entre los valores asumidos por cada variable en el par de observaciones en cuestión,

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}$$

Esta distancia varía con la escala de las coordenadas, por ello puede ser completamente distorsionada por un simple cambio de unidad de medida de las variables en estudio. Como ejemplo véase la siguiente tabla:

Tabla 59: Ejemplo del peso y altura de tres personas.

Persona	Peso (en libras)	Altura (en pies)
A	160	5.5
B	163	6.2
C	165	6.0

Las distancias Euclídeas son: d_{AB}=3.08; d_{AC}=5.02; d_{BC}=2.01. Sin embargo, si la altura se mide en pulgadas, las distancias serán: d_{AB}=8.92; d_{AC}=7.81; d_{BC}=3.12. En este último caso, la persona A está más cercana de la C que de la B, mientras que en el caso anterior sucede lo contrario. La falta de invariancia respecto de la escala sugiere *estandarizar* los datos (dividiendo por la desviación estándar) antes de calcular las distancias Euclídeas. En este sentido, las distancias relativas (distancias entre rangos) no varían.

La distancia **Manhattan** es la suma de los valores absolutos de las diferencias entre cada par de coordenadas que define la observación p-dimensional. La métrica de Manhattan suele ser utilizada con datos en escala ordinal y por intervalos mientras que la métrica Euclídea es usada con datos continuos. La ausencia de observaciones en una o más variables acarrea problemas a la hora de calcular distancias. En la práctica observaciones multivariadas con algunos valores perdidos suelen ser descartadas completamente. Si forman parte del análisis se pueden calcular *distancias promedio* obtenidas dividiendo el valor de distancia obtenidos entre dos observaciones por el número de variables con realizaciones en ambas observaciones. De esta manera la comparación relativa de magnitudes de distancia no se ve sesgada si las distancias han sido calculadas a partir de distintos números de pares coordenados. En este sentido, InfoStat provee las distancias: **Euclídea promedio** y **Manhattan promedio**. Cuando se dispone de variables no-métricas o mezclas de variables

métricas y no métricas, las distancias geométricas clásicas presentadas anteriormente para medir proximidad de los objetos no son recomendadas. En tales situaciones conviene utilizar medidas de distancia que no exigen que se cumpla la desigualdad triangular (Spath, 1980).

En muchas aplicaciones, sobre cada unidad de muestreo se mide la presencia o ausencia de *p* características y las similaridades entre las unidades u observaciones deben ser construidas desde la información sobre presencia/ausencia. Varias medidas de similitud para datos dicotómicos o binarios se derivan desde tablas de contingencia de dos conjuntos de datos indicando presencia o ausencia. Suponga que a partir de las características observadas en cada uno de dos sujetos o unidades de muestreo, se construye la siguiente tabla:

Tabla 60: Ejemplo organización de datos binarios.

	Sujeto 1	
Sujeto 2	Presencia	Ausencia
Presencia	a	b
Ausencia	c	d

Luego, **a** representa el número de características presentes en ambos sujetos en la muestra, **d** representa el número de características ausentes en ambos, y **b** y **c** son el número de características presentes en un sujeto pero ausente en otro. Dicho de otra manera, **a** y **d** representan coincidencias o pares asociados positivamente (1,1) o negativamente (0,0). Por el contrario b y c representan pares no asociados. Las distancias basadas en el índice de Jaccard, el coeficiente de apareamiento simple (simple matching), el coeficiente de apareamiento positivo, el índice de Rogers y Tanimoto, y el índice de Anderberg, entre otros, son utilizadas como métrica de similitud obtenida desde este tipo de información.

Existen diversas medidas de asociación a los fines de ponderar el significado de la presencia conjunta o la ausencia conjunta de características según corresponda en cada caso. En algunas circunstancias la ausencia conjunta, por ejemplo, puede no implicar similitud entre dos observaciones. Por ejemplo, en estudios de ecología de vegetación, la presencia de dos plantas raras sobre dos parcelas de muestreo puede implicar que ambas parcelas son parecidas, mientras que su ausencia conjunta sobre otras dos parcelas probablemente no diga nada acerca de la similitud de ellas. Las mayores diferencias entre estos coeficientes se deben a: 1) si las asociaciones negativas están incorporadas en la medida; 2) si los pares asociados tienen igual peso que los pares no asociados y 3) si los pares no asociados tienen distinto peso que los pares asociados. La siguiente tabla muestra el cálculo de distintos coeficientes de asociación o semejanza provistos por InfoStat:

Tabla 61: Coeficientes de similitud.

1. Single matching	$(a+d)/(a+b+c+d)$
2. Positive matching	$a/(a+b+c+d)$
3. Jaccard	$a/(a+b+c)$
4. Anderberg	$a/[a+2(b+c)]$
5. Roger y Tanimoto	$(a+d)/(a+d+2(b+c))$

6. Ochiai	$a/\sqrt{(a+b)(a+c)}$
7. Dice	$2a/(2a+b+c)$
8. Kulczynski 1	$a/(b+c)$
9. Kulczynski 2	$0.5\{[a/(a+c)]+[a/(a+b)]\}$
10. Braun-Blanquet	a/max[(a+b),(a+c)]
11. Hamman	$[(a+d)-(b+c)]/(a+b+c+d)$
12. Sokal y Sneath 1	$(a+d)/(a+d+0.5(b+c))$
13. Sokal y Sneath 2	$0.25\{[a/(a+b)]+[a/(a+c)]+[d/(d+b)]+[d/(d+c)]\}$
14. Sokal y Sneath 3	$(a.d)/\sqrt{[(a+b).(a+c).(d+b).(d+c)]}$
15. Coeficiente Phi	$[(a.d)-(c.b)]/\sqrt{[(a+b).(a+c).(b+d).(c+d)]}$
16. Yule y Kendall	$[(a.d)-(b.c)]/[(a.d)+(b.c)]$

La medida **simple matching** (apareamiento simple), y su pondera de la misma forma las coincidencias establecidas como **a** que aquellas contabilizadas por **d**.

La medida **positive matching** (apareamiento positivo), es útil cuando la presencia simultánea es más importante para cuantificar la similitud que la ausencia simultánea.

El **índice de Jaccard**, es útil si se desea enfatizar la posesión de los atributos, es decir las situaciones (1,1), (1,0) y (0,1).

El **índice de Anderberg** se caracteriza por dar mayor peso a los pares no asociados (1,0) y (0,1).

El **coeficiente de Rogers y Tanimoto** le da doble peso a los pares no asociados, pero considera el (0,0) en su cálculo.

Cuando se trabaja con variables ordinales o cuando el interés radica en agrupar variables más que observaciones, se recomiendan medidas de similitud que se basan en coeficientes de correlación muestral (Pearson y Spearman). Ambos coeficientes de correlación toman valores entre –1 y 1. El signo indica el sentido de la correlación y el valor absoluto mide la fuerza de la correlación. El **coeficiente de Spearman** puede ser visualizado como una versión no-paramétrica del **coeficiente de Pearson**, ya que los datos son transformados en rangos antes de calcular la correlación (ver Correlación). El coeficiente no-paramétrico **Tau de Kendal** trabaja con los rangos de las variables ordenados, asume también valores entre –1 y 1. InfoStat permite implementar las siguientes medidas de distancias basadas en medidas de similitud **1-S, sqrt(1-S), -log(S), 1/S-1, sqrt(2(1-S)), 1-(S+1)/2, 1-abs(S) y 1/(S+1),** donde S es la similitud a partir de la cual se quiere obtener una distancia, *sqrt* es la raíz cuadrada, *abs* es la función valor absoluto y *log* es el logaritmo (ver Tabla 9). Por otra parte, cabe destacar que el coeficiente de correlación está relacionado con el estadístico Chi cuadrado ($r^2 = \chi^2 / n$) para la prueba de independencia de dos variables categóricas. Para *n* fijo, una alta correlación (similitud) es consistente con falta de independencia (Johnson y

Wichern, 1998). InfoStat provee también la distancia **Chi cuadrado**, la cual es obtenida a partir del estadístico $\chi 2$ clásico para tablas de contingencia como una medida de distancia de cada celda con respecto a su valor esperado.

Tabla 62: Funciones para la obtención de medidas de distancia a partir de índices de similitud.

Función	Rango para S_{ij}	Rango para d_{ij}
1. $d_{ij} = 1 - S_{ij}$	$[0,1]$	$[0,1]$
2. $d_{ij} = \sqrt{1 - S_{ij}}$	$[0,1]$	$[0,1]$
3. $d_{ij} = -\log S_{ij}$	$(0,1]$	$[0,\infty)$
4. $d_{ij} = 1/S_{ij} - 1$	$(0,1]$	$[0,\infty)$
5. $d_{ij} = \sqrt{2(1 - S_{ij})}$	$[0,1]$	$[0,\sqrt{2}]$
6. $d_{ij} = 1 - (S_{ij} + 1)/2$	$[-1,1]$	$[0,1]$
7. $d_{ij} = \sqrt{1 - (S_{ij} + 1)/2}$	$[-1,1]$	$[0,1]$

Análisis de correspondencias

Menú ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DE CORRESPONDENCIAS permite realizar análisis de correspondencias simple y múltiple sobre tablas de datos categorizadas conteniendo variables.

En el **Selector de variables** se pueden especificar los **Criterios de clasificación** y si es necesario las **Frecuencias**. Al **Aceptar** aparecerá una ventana de diálogo en la que se pueden elegir las siguientes opciones: **Frecuencias absolutas**, **Perfiles fila**, **Perfiles columna**, **Frecuencias relativas al total**, **Frecuencias esperadas para el estadístico Chi cuadrado**, **Desviaciones respecto de lo esperado bajo independencia**, **Contribuciones individuales al estadístico chi-cuadrado**, **Valores singulares**, **Coordenadas fila**, **Coordenadas columna**, **Biplot**. Se encuentran también las opciones **Frecuencias relativas como porcentajes** y **Extraer 2 ejes**, siendo ambas modificables.

Ejemplo 39: *El AC simple se llevó a cabo sobre un estudio que aborda la caracterización de mujeres con problemas relacionados con el alcohol desde características sociodemográficas y psicológicas. Si bien se relevaron un conjunto de variables categorizadas tales como edad, ocupación, estado civil, motivo de consulta y diagnóstico del paciente al entrar al centro de rehabilitación. Se usó AC simple para estudiar la asociación entre motivo de consulta y edad. Los datos (gentileza de Yolanda Prados y Graciela Diosque, Facultad de Psicología, U.N.C), se encuentran en el archivo Alcoholismo.*

Se presenta el biplot obtenido al realizar el AC simple de las variables "motivo de consulta" y "edad". Las modalidades de la variable motivo de consulta fueron: C-Far (uso de fármacos), C-Sus (uso de sustancias que generan adición), C-Der (derivados de otros consultorios), C-Des (deseos de dejar de beber), C-Alc (consumo de alcohol), C-EsA

(estado de ánimo), C-Vio (violencia familiar), C-Fis (síntomas físicos). Para la variable "edad" las modalidades fueron: Jov (menos de 30 años), Med (entre 30 y 50 años) y May (más de 50 años). Para obtener el biplot se usaron los siguientes comandos:

Menú ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DE CORRESPONDENCIA. En **Criterios de clasificación** se eligieron "motivo de consulta" y "edad". En la siguiente ventana se dejaron las opciones por defecto: **Frecuencias absolutas**, **Perfiles fila**, **Perfiles columna**, **Contribuciones individuales al estadístico Chi cuadrado**, **Valores singulares**, **Coordenadas fila**, **Coordenadas columna**, **Biplot** y **Extraer 2 ejes**.

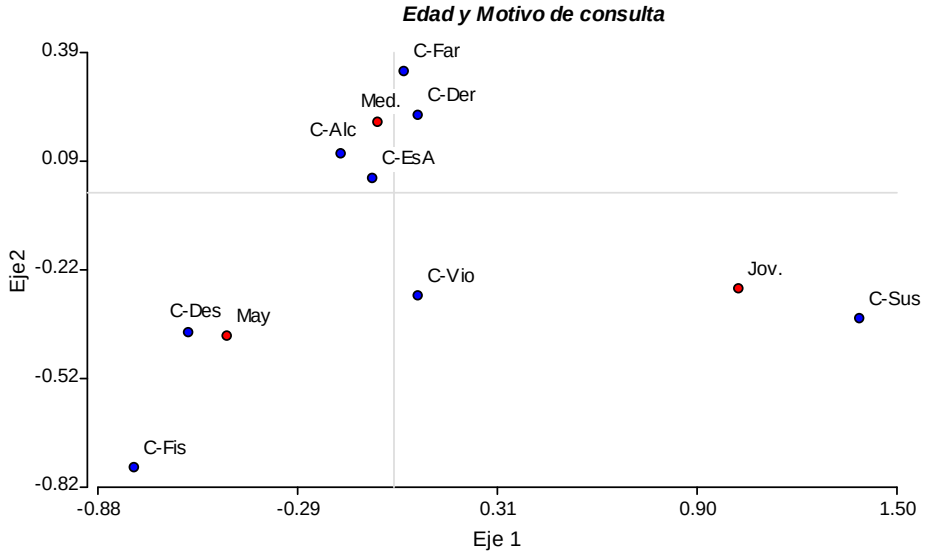


Figura 31: Biplot. Archivo Alcoholismo.

En la figura se visualizan las dos primeras dimensiones del AC simple de la tabla de contingencia correspondiente al cruce de las variables "edad" y "motivo de consulta". El gráfico sugiere, en su primer eje (con una inercia de 73.99%), que las jóvenes (menores de 30 años) consultaban mayoritariamente por consumo de sustancias (C-Sus) y que los mayores de 50 años consultaban por deseos de dejar de beber (C-Des) y por síntomas físicos (C-Fis). Las mujeres de edad media citaban al consumo de alcohol, el estado de ánimo, consumo de psicofármacos y derivación como los principales motivos de consulta. Los puntos representando modalidades de una misma variable respuesta pueden ser automáticamente conectados en InfoStat. En la ventana **Resultados**, se obtuvo una tabla como la siguiente:

Tabla 63: Análisis de correspondencias simple. Archivo Alcoholismo.

Frecuencias absolutas
En columnas: MOTIVO
En filas: EDAD

	C-Alc	C-Der	C-Des	C-EsA	C-Far	C-Fis	C-Sus	C-Vio	Total
Jov.	2	3	0	2	1	0	5	2	15

May	6	3	5	4	1	2	0	3	24
Med.	19	17	5	12	8	1	3	5	70
Total	27	23	10	18	10	3	8	10	109

Perfiles fila (frecuencias relativas por filas)
En columnas: MOTIVO
En filas: EDAD

	C-Alc	C-Der	C-Des	C-EsA	C-Far	C-Fis	C-Sus	C-Vio	Total
Jov.	0.13	0.20	0.00	0.13	0.07	0.00	0.33	0.13	1.00
May	0.25	0.13	0.21	0.17	0.04	0.08	0.00	0.13	1.00
Med.	0.27	0.24	0.07	0.17	0.11	0.01	0.04	0.07	1.00
Total	0.25	0.21	0.09	0.17	0.09	0.03	0.07	0.09	1.00

Perfiles columna (frecuencias relativas por columnas)
En columnas: MOTIVO
En filas: EDAD

	C-Alc	C-Der	C-Des	C-EsA	C-Far	C-Fis	C-Sus	C-Vio	Total
Jov.	0.07	0.13	0.00	0.11	0.10	0.00	0.63	0.20	0.14
May	0.22	0.13	0.50	0.22	0.10	0.67	0.00	0.30	0.22
Med.	0.70	0.74	0.50	0.67	0.80	0.33	0.38	0.50	0.64
Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Contribuciones por celda al estadístico chi-cuadrado
En columnas: MOTIVO
En filas: EDAD

	C-Alc	C-Der	C-Des	C-EsA	C-Far	C-Fis	C-Sus	C-Vio	Total
Jov.	0.79	0.01	1.38	0.09	0.10	0.41	13.81	0.28	16.88
May	5.1E-04	0.84	3.56	3.4E-04	0.66	2.72	1.76	0.29	9.82
Med.	0.16	0.34	0.31	0.02	0.39	0.45	0.89	0.31	2.86
Total	0.95	1.19	5.25	0.11	1.15	3.57	16.46	0.89	29.56

Contribución a la Chi cuadrado

	Autovalor	Inercias	Chi-Cuadrado	(%)	% acumulado
1	0.45	0.20	21.87	73.99	73.99
2	0.27	0.07	7.69	26.01	100.00

Nociones teóricas sobre el análisis de correspondencias

El análisis de correspondencias (AC) es una técnica exploratoria que permite representar gráficamente filas y columnas de una tabla de contingencia (Greenacre, 1984, 1988,1994; Lebart *et al*, 1984). En Psicología suelen referirse a esta técnica como escalamiento dual; en Ecología ha sido muy usada para ordenamiento de datos discretos de vegetación (presencia-ausencia de una serie de especies en cada parcela observada a lo largo de un gradiente ambiental). La técnica de AC también constituye una herramienta de principal importancia para el análisis de datos textuales donde se construyen tablas de contingencia relacionando el uso de varios vocablos entre distintos textos de discurso. El AC puede ser interpretado como una técnica complementaria y a veces suplementaria del uso de modelos log-lineales para el estudio analítico de las relaciones contenidas en tablas de contingencia. El AC permite explorar gráficamente estas relaciones.

En el AC se representan las filas y las columnas de una tabla a dos vías de variables categorizadas, como puntos en un espacio Euclídeo de baja dimensión (generalmente bidimensional). El propósito de su uso es similar al del análisis de componentes principales

para datos continuos, diferenciándose de este por el hecho de que el AC opera sobre la matriz de desviaciones Chi cuadrado en lugar de usar una matriz de varianzas y covarianzas.

Las filas de la tabla de contingencia pueden ser vistas como puntos con coordenadas dadas por las columnas de la tabla. Los perfiles filas son construidos a partir de la división de la frecuencia observada en cada celda por el correspondiente total de fila. A cada punto fila se le asigna un peso a través de la división del total de fila por el gran total de la tabla. Los perfiles columnas se definen de manera equivalente. El AC determina, a través de la descomposición por valor singular de la matriz de desviaciones Chi cuadrado de proporciones filas y columnas bajo la hipótesis de independencia entre filas y columnas, un subespacio óptimo para la representación de los perfiles files y columnas ponderados por sus respectivos pesos.

Cuando el AC es realizado sobre una única tabla de dos vías, se denomina Análisis de Correspondencia Simple (ACS). Este análisis permite graficar observaciones bivariadas en planos e identificar las asociaciones de mayor peso entre las modalidades de dos variables cualitativas. El Análisis de Correspondencia Múltiple (ACM) permite explorar tablas multidimensionales. Las observaciones multivariadas se grafican en planos para así poder identificar las asociaciones de mayor peso entre las modalidades de varias variables cualitativas. Para este último enfoque se utilizan las conocidas tablas *Burt* que contienen los niveles o modalidades de cada variable categorizada tanto en las filas como en las columnas de la tabla y por tanto contienen todas las clasificaciones cruzadas a dos vías de las variables originales (Greenacre, 1984).

El AC opera sobre la matriz de desviaciones Chi cuadrado, en lugar de usar la matriz de varianzas y covarianzas como lo hace el análisis de componentes principales. Este método mide cuales son las combinaciones de modalidades que tienen más *inercia* (que más contribuyen a rechazar la hipótesis de independencia entre las dos variables). Ellas son las modalidades de la periferia o modalidades que se alejan del centro del plano. Como el análisis no es realizado sobre las frecuencias absolutas sino sobre las proporciones de la tabla de contingencia, comúnmente se utiliza el término *inercia* para denotar la información Chi cuadrado en la tabla (inercia es el valor Chi cuadrado dividido por el gran total de la tabla).

Para la matriz de desviaciones por celda se obtiene un conjunto de autovectores y autovalores que son usados para construir un subespacio óptimo para la representación de los perfiles filas y columnas ponderados por sus respectivos pesos. Los ejes son extraídos en relación a la desviación Chi cuadrado explicada por cada uno. El primer eje principal se asocia a la más alta contribución sobre el estadístico Chi cuadrado de la tabla de contingencia. Los primeros d ejes definen el espacio d-dimensional óptimo, con $d=min(I-1,J-1)$ con I=número de filas, J=número de columnas. La proporción de la inercia total explicada por cada eje es usada como criterio de selección del número de ejes necesarios para la representación.

Similar al análisis de componentes principales, los resultados pueden ser representados en un biplot para graficar los puntos filas y columnas en el mismo espacio (Greenacre y Hastie, 1987). Las distancias entre puntos filas miden la discrepancia entre perfiles filas. Los puntos filas muy cercanos en el gráfico, tienen similar perfil fila. Distancias desde el origen indican

la discrepancia entre los perfiles filas y el centroide fila o la distribución fila marginal. El mismo tipo de interpretación puede realizarse sobre los perfiles columnas. Las distancias entre puntos filas y columnas carecen de sentido, pero puntos filas y columnas que caen en la misma dirección respecto al origen se encuentran positivamente correlacionados, mientras que aquellos que caen en direcciones opuestas se encuentras negativamente correlacionados. Las direcciones pueden cambiar si se grafican otras dimensiones, por lo que es importante realizar el análisis sobre un espacio con alta inercia.

Análisis de coordenadas principales

Menú ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DE COORDENADAS PRINCIPALES permite analizar la interdependencia entre variables categóricas y encontrar una representación grafica de los *n* individuos tal que se refleje la distancia entre ellos. Estas distancias pueden ser calculadas a partir de la estructura de similitudes definida por la matriz de similitudes S. A diferencia del análisis de componentes principales que requiere variables cuantitativas en análisis de coordenadas principales (ACoorP) se puede hacer con cualquier tipo de variables, incluso con mezcla de variables.

En la ventana **Análisis de coordenadas principales**, se deben indicar las variables respuesta y las de clasificación en caso que existan (opcional). En la solapa **Medidas resumen** hay opciones para guardar las coordenadas obtenidas (**Ejes**) según el número de coordenadas que se indique. Si se guardan coordenadas principales se adicionarán como nuevas columnas a la tabla activa. Estas coordenadas pueden ser utilizadas posteriormente para realizar gráficos de dispersión de las observaciones. Se puede pedir la estandarización de cada variable antes de comenzar el análisis (**Estandarizar datos)**, la visualización de la matriz de distancias (**Mostrar matriz de distancias**) sobre la que se realiza el análisis y el **Arbol de recorrido mínimo (ARM)**. Los datos se pueden **canalizar por filas o por columnas** y además se pueden seleccionar entre dos funciones de distancia: $\mathbf{M_{ij}=-0.5*D_{ij}*D_{ij}}$ **o** $\mathbf{M_{ij}=1/(1+D_{ij})}$ frecuentemente usadas para convertir similitudes a distancias, en el marco del ACoorP**.** En caso de que ese haya indicado un criterio clasificación en la solapa **Medidas de resumen**, InfoStat permite escoger entre medidas de posición como la **media**, **mediana**, **mínimo**, **máximo** y de dispersión como la **varianza** y desviación estándar (**desvío**) como estadísticos para resumir la información de cada variable en cada conjunto de registros indexado por el criterio (opcional).

Ejemplo 40: En un estudio que tuvo como objetivo estudiar los alimentos que se utilizan como fuentes proteicas, en las dietas de los habitantes de países europeos, se registraron los alimentos consumidos. Después de un primer análisis que indicó que la principal causa de variabilidad en los hábitos era el consumo de productos cárnicos, se desea analizar la distancia entre los países calculadas a partir de 4 variables que se relaciones con fuentes proteicas de origen cárnico. Los datos se encuentran en el archivo Proteínas.

Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ ANÁLISIS DE COORDENADAS PRINCIPALES. En la ventana **Análisis de coordenadas principales** seleccionar “Carne Vacuna”, “Carne Cerdo”, “Huevo” y “Leche” como **Variables** y “País” como **Criterio de clasificación.** En la ventana **Análisis de coordenadas principales**, se

activó **Guardar 2 ejes o coordenadas principales.** También se activó **Estandarizar** para realizar el cálculo de distancias (dada la naturaleza de la variable se seleccionó distancias Euclídeas) sobre la matriz de datos estandarizados y se activó la opción **ARM** a partir de la cual se obtuvo el siguiente plano que explica el 82% de la variación total y el **ARM** que permite identificar cuales son los países mas cercanos en cuanto al habito de consume si se consideran estas 4 variables.

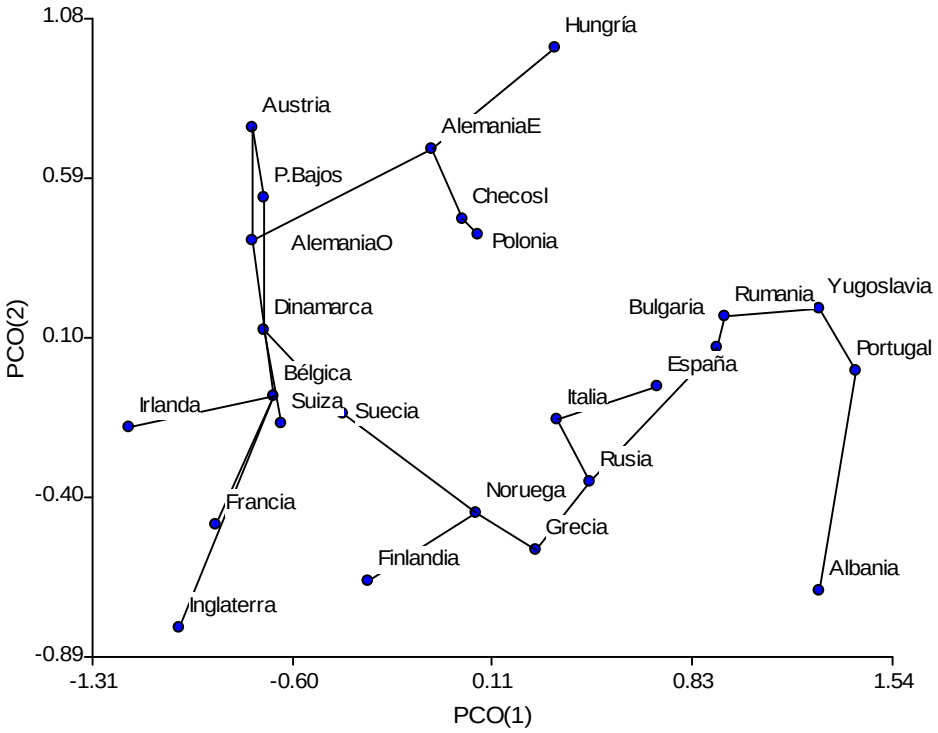


Figura 32: Proyección de observaciones multivariadas (países) en el espacio de las dos primeras coordenadas principales y ARM

Nociones teóricas sobre el análisis de coordenadas principales

La técnica de reducción conocida como análisis de coordenadas principales (ACoorP) es una forma de escalamiento multidimensional clásico o métrico. La técnica de escalamiento multidimensional explora las similitudes (o distancias) entre observaciones y permite mostrarlas de manera gráfica. Es una técnica útil para mostrar distancias entre datos para los cuales las medidas Euclídeas no son apropiadas o se desea, por alguna otra razón, usar una medida de distancia alternativa expresada como función de un índice de asociación.

El objetivo de la técnica es mostrar las relaciones entre observaciones, representadas por distancias o similitudes, en un plano tal que las distancias verdaderas sean preservadas tanto como sea posible. La técnica de escalamiento multidimensional usa la matriz de distancia o

de similitudes para construir la configuración de puntos en el plano. El AcoorP opera sobre una matriz de similitudes doblemente centrada derivada de la matriz de similitud (también de distancia) como sigue: $C_{ij} = A_{ij} - \overline{A}_{i.} - \overline{A}_{.j} + \overline{A}_{..}$. Se realiza la descomposición espectral de la nueva matriz, **C**, y se obtiene la solución o coordenadas principales haciendo, $\mathbf{Z} = \mathbf{E}\mathbf{D}^{1/2}$ si $\mathbf{C} = \mathbf{E}\mathbf{D}\mathbf{E}'$ representa la descomposición espectral de **C**. Los autovalores de la descomposición son los elementos diagonales de la matriz **D**, cada uno indica la cantidad de variabilidad explicada en la dimensión dada por el autovector correspondiente de la matriz **E**.

Árboles de clasificación y árboles de regresión

Menú ⇒ ANÁLISIS MULTIVARIADO ⇒ ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN permite clasificar observaciones multivariadas en forma de árboles de decisión.

En el Selector de variables se pueden especificar la Variable dependiente y las Variables regresoras. Al Aceptar aparecerá una ventana de diálogo en la que se pueden elegir la Medida de heterogeneidad dentro de los nodos (H), el Mínimo tamaño del nodo para continuar la partición (*n*) y el Umbral de heterogeneidad dentro del grupo para terminar.

InfoStat provee dos medidas de heterogeneidad dentro de nodos (H): la **Deviance**, recomendada cuando la variable dependiente es una variable de clasificación y la **Suma de cuadrados** de los valores de la variable dependiente dentro de cada nodo, medida usualmente seleccionada para variables continuas.

Se presenta a continuación un árbol de clasificación para los datos del archivo *Iris*. Para realizarlo elegir: Menú ESTADÍSTICAS ⇒ ANÁLISIS MULTIVARIADO ⇒ ÁRBOLES DE CLASIFICACIÓN (REGRESIÓN TREES). En la ventana correspondiente designar a "PetalLen", "PetalWid", "SepalLen" y "SepalWid" (en ese orden) como **Variables regresoras** y a "Especie" como **Variable dependiente**. En la siguiente ventana de diálogo elegir la **Deviance** como **Medida de heterogeneidad** ya que la variable dependiente es una variable de clasificación (especie) y dejar las otras opciones propuestas por defecto para obtener el siguiente árbol de clasificación:

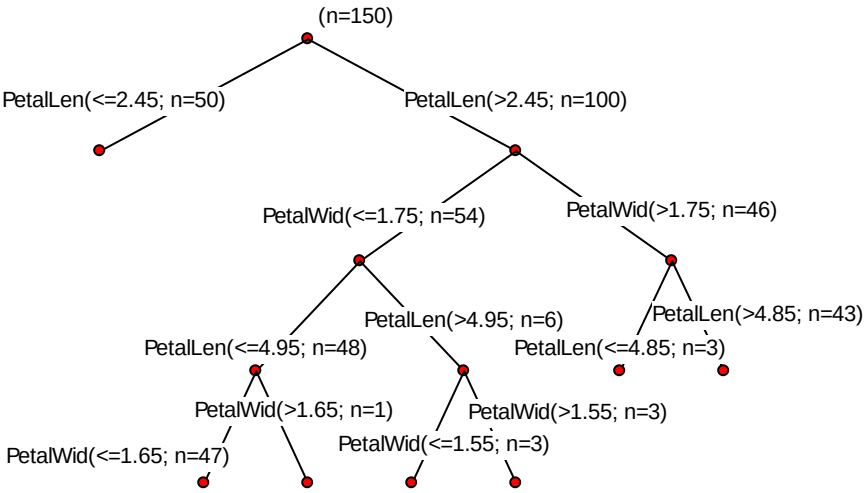


Figura 33: Árbol de clasificación. Archivo Iris.

Como puede observarse la primera separación se realiza en base a valores de la variable largo del pétalo menores e iguales a 2.45 (50 individuos) y los mayores a 2.45 (100 individuos). De esta rama separa en base al ancho del pétalo los menores e iguales a 1.75 (54 individuos) y los mayores de 1.75 (46 individuos) y así el proceso continúa.

Nota: En este archivo de datos se produce un hecho particular: las dos variables ancho del pétalo y largo del pétalo separan el primer nodo de la misma forma ya que la especie 1 se caracteriza por tener corolas más grandes que las otras dos especies en ambas dimensiones del pétalo (largo y ancho). Situaciones como estas son resueltas por InfoStat eligiendo entre aquellas variables con idéntico potencial de discriminación, la que se encuentra primero en la lista de variables regresoras.

Se recomienda ingresar las variables según el estadístico F que puede obtenerse con un análisis de la varianza univariado. En este caso el estadístico F más grande corresponde al largo del pétalo. Es por ello que se la dispone como la primera variable regresora.

En **Resultados** se obtiene automáticamente el árbol y la historia de formación de nodos, incluyendo los valores del estadístico usado para la clasificación y los puntos críticos o valores de las variables asociadas a cada nodo.

Nociones teóricas sobre árboles de clasificación y regresión

Los modelos basados en árboles de regresión y/o clasificación constituyen una alternativa a los modelos lineales aditivos para los problemas de regresión y para modelos logísticos aditivos en problemas de clasificación. Estos modelos están pensados para captar comportamientos no aditivos; los modelos lineales estándar no permiten interacciones entre variables a menos que se especifique una forma multiplicativa. En ciertas aplicaciones, especialmente cuando el grupo de predictores contiene una mezcla de variables numéricas y factores, los modelos basados en árboles son más fáciles para interpretar y discutir que los

modelos lineales. Se denominan modelos de árbol porque el método original de presentar los resultados es en forma de árbol binario. Cuando la variable dependiente es continua se conforman *árboles de regresión* y cuando es de clasificación se generan *árboles de clasificación*.

Un árbol de regresión o de clasificación es un conjunto de muchas reglas determinadas por un procedimiento de ajuste por particiones binarias recursivas, donde un conjunto de datos es sucesivamente particionado. Esta técnica está relacionada con los conglomerados divisivos. Inicialmente todos los objetos son considerados como pertenecientes al mismo grupo. El grupo se separa en dos subgrupos a partir de una de las *variables regresoras* de manera tal que la heterogeneidad, a nivel de la *variable dependiente*, sea mínima de acuerdo a la medida de heterogeneidad seleccionada. Los dos subgrupos (nodos) formados se separaran nuevamente si: 1) hay suficiente heterogeneidad para producir una partición de observaciones y/o 2) el tamaño del nodo es superior al mínimo establecido para continuar el algoritmo. El proceso se detiene cuando no se cumple una de estas condiciones. En cada instancia de separación el algoritmo analiza todas las *variables regresoras* y selecciona, para realizar la partición, aquella que permite conformar grupos más homogéneos dentro y más heterogéneos entre ellos.

Biplot y árbol de mínimo recorrido

Los gráficos de dispersión son usados para visualizar directamente las observaciones o las variables, las relaciones en otra dimensión son sólo implícitas. Los gráficos Biplots propuestos por Gabriel (1971,1981), muestran las observaciones y las variables en el mismo gráfico, de forma tal que se pueden hacer interpretaciones sobre las relaciones conjuntas. El prefijo "bi" en el nombre *biplot* refleja la característica de que tanto observaciones como variables, son representadas en el mismo gráfico.

En los biplots, las observaciones son generalmente graficadas como puntos. La configuración de los puntos es obtenida a partir de combinaciones lineales de las variables originales. Las variables son graficadas como vectores desde el origen. Los ángulos entre las variables representan la correlación entre las variables.

Las dimensiones seleccionadas para el biplot son aquellas que mejor explican la variabilidad de los datos originales. Para encontrar los ejes óptimos para la graficación de observaciones y variables en un espacio común se utiliza la idea de que cualquier matriz de datos $n \times p$, puede ser representada aproximadamente en d dimensiones como el producto de dos matrices, $\mathbf{A}$ ($n \times d$) y $\mathbf{B}$ ($p \times d$), y d es el rango de la matriz original, así $\mathbf{AB}'$ aproxima la matriz original. Debido a que $\mathbf{A}$ y $\mathbf{B}$ tienen una base común de d vectores, es posible mostrar las filas y las columnas de la matriz original sobre el mismo gráfico con varias condiciones de optimalidad y con la posibilidad de realizar interpretaciones sobre las distancias entre puntos.

Los biplots pueden considerarse como una técnica de reducción de dimensión ya que las filas de $\mathbf{A}$ representan las observaciones en un espacio de menor dimensión (puntos fila) y las columnas de $\mathbf{B}'$ representan las variables (puntos columnas) en ese mismo espacio. Si

la descomposición por valor singular de X es $\mathbf{X} = \mathbf{U}\mathbf{D}_{\Gamma}V'$, donde U es $n \times p$ con columnas ortogonales, V es una matriz ortogonal $p \times p$ y $\mathbf{D}_{\Gamma}$ una matriz diagonal $p \times p$ de valores singulares, las matrices **A** y **B** pueden ser expresadas como

$$A = UD_{\Gamma}^{\alpha} \text{ y } B = VD_{\Gamma}^{1-\alpha}$$

donde α es usualmente igual a 0, ½, ó 1 para proveer condiciones de optimalidad en el gráfico (Gower y Digby, 1981). Los Biplots son gráficos de dispersión de los $n+p$ vectores de **A** y **B** en un mismo espacio d-dimensional. Comúnmente se presentan gráficos bidimensionales seleccionado las dos componentes en **U** y **V** asociados con los dos valores singulares más altos en **D**. Estos gráficos son aproximaciones de la matriz original a menos que toda la variabilidad sea explicada por los dos primeros ejes.

En los biplots la distancia entre símbolos representando observaciones y símbolos representando variables no tiene interpretación, pero las direcciones de los símbolos desde el origen si pueden ser interpretadas. Las observaciones (puntos filas) que se grafican en una misma dirección que una variable (punto columna) podría tener valores relativamente altos para esa variable y valores bajos en variables o puntos columnas que se grafican en dirección opuesta. Dependiendo de las condiciones de optimalidad especificadas, las distancias entre los puntos filas o entre los puntos columnas pueden ser estadísticamente interpretadas, los ángulos entre los vectores que representan las variables pueden ser interpretados en términos de las correlaciones entre variables y las longitudes de los rayos pueden hacerse proporcionales a las desviaciones estándar. Cuando las longitudes de los vectores son similares el gráfico sugiere contribuciones similares de cada variable en la representación realizada.

InfoStat permite obtener biplots desde la matriz de datos y para otras matrices en el marco de distintos procedimientos de análisis multivariado como es el caso de los biplots solicitados desde el ACP y desde el AD. Cuando los Biplots son pedidos desde el menú principal de análisis multivariado, el usuario puede modificar el valor de α . En otros casos, InfoStat asigna el valor ½ al coeficiente α , en este caso el biplot es conocido como biplot simétrico. Con α igual a 0 y 1 se obtienen mejores representaciones del espacio columna y del espacio fila, respectivamente.

Si se activa la casilla **ARM**, el usuario puede visualizar sobre el mismo gráfico biplot solicitado un árbol de mínimo recorrido. Los árboles de recorrido se construyen uniendo puntos que representan observaciones multivariadas y que se proyectan en un plano como resultado de alguna técnica de reducción de dimensión. Los puntos son conectados con segmentos de líneas rectas tal que todos los puntos quedan unidos directa o indirectamente y no hay *loops* (Gower y Ross, 1969). El árbol de mínimo recorrido es un árbol de recorrido con segmentos conectados de tal manera que la suma de las longitudes de todos los segmentos es mínima. El árbol de mínimo recorrido puede calcularse a partir de la matriz de distancia de las observaciones multivariadas en el espacio p-dimensional en el que viven o a partir de matrices de distancia en espacios de menor dimensión. Cuando puntos p-dimensionales (con p>2) son conectados, en el plano, en función de su distancia en el espacio original, el árbol de mínimo recorrido puede proveer información sobre similitudes de las observaciones en otras dimensiones no directamente representadas en el plano. Por

ejemplo, algunos puntos que se encuentran muy cerca en el espacio bidimensional podrían estar, en su espacio original, más lejos de lo que aparentan en el plano. Los árboles de mínimo recorrido conceptualmente se ligan al algoritmo de agrupamiento conocido como encadenamiento simple y en ese sentido son usados no solo para representación gráfica sino también para formar conglomerados de puntos. En la ventana de **Herramienta Gráficas**, InfoStat también presenta la opción **ARM**, ligada a cualquier serie graficada, pero en este caso el árbol generado es obtenido a partir de la matriz de distancias de los puntos bi-dimensionales que se están graficando. Esta opción solo conecta los puntos en función de las distancias que el usuario esta visualizando en el plano.

Procrustes generalizado

Las configuraciones geométricas obtenidas mediante escalamientos multidimensionales, coordenadas principales u otras técnicas similares, ofrecen una de las maneras más clásicas de representar la estructura y relación empírica de un conjunto de elementos o individuos a los cuales se les ha observado simultáneamente varios atributos. En muchos casos, la orientación de las dimensiones es arbitraria, y cuando se han obtenido varias configuraciones sobre la misma muestra de elementos ya sea porque se realizaron en diferentes momentos o por que participaron distintos observadores o porque se utilizaron diversas técnicas para realizar el ordenamiento, se requiere de una técnica para analizar la congruencia de dichas configuraciones. El análisis de procrustes es utilizado con tal fin.

Bramardi (2001) comenta que la palabra procrustes fue utilizada por primera vez para describir la armonización o adecuación de configuraciones, en referencia a un término de origen griego que significa “martillear” y hace alusión a un posadero mitologico, quien estiraba o recortaba a los huéspedes sus extremidades de modo que coincidieran con la cama en la que se acostaban. Inicialmente el análisis de procrustes fue utilizado para adecuar o ajustar una configuración a otra ya representada. Se describió la adecuación de configuraciones como una transformación en que una matriz es rotada y constreñida según especificaciones de una matriz establecida a la que se denomina matriz objetivo. La matriz transformada debe coincidir tanto como sea posible con la matriz objetivo, esto es lo que se conoce como transformación procrusteana. El método propuesto es restrictivo para matrices con igual número de columnas y de rango completo y se basa en un criterio de mínimos cuadrados que minimizan las distancias entre los puntos análogos en la configuración final. Bajo el criterio de rotar una matriz para ajustarla a otra, es posible rotar varias matrices a una matriz centroide común. Esto es lo que se conoce con el nombre de análisis de procrustes generalizado. Gower (1975) describe la matriz centroide como una representación de configuración promedio de consenso e incluye la traslación y el escalamiento de las matrices previa estandarización de las mismas en su análisis, proponiendo una técnica de cálculo que culmina con un formato de análisis de la varianza.

La técnica de cálculo par el análisis de procrustes generalizado desarrollado por Gower propone la armonización de las configuraciones individuales a través de una serie de pasos iterativos por transformación de éstas.Los sucesivos pasos o transformaciones que se realizan en un análisis de procrustes generalizado incluyen normalización, rotación, reflexión y escalamiento de los datos bajo dos criterios: (1) que se mantengan las distancias

entre los individuos de las configuraciones individuales, y (2) que se minimice la suma de cuadrados entre puntos análogos, es decir correspondientes al mismo elemento, y su centroide. La configuración de consenso se obtiene como la media de todas esas configuraciones individuales tranformadas.

En términos matriciales, si cada matriz individual está representada por $\mathbf{X}_i$, (i=1, 2, ...,*m*) con *n* filas y *p* columnas donde la j-ésima fila da las coordenadas de un punto (individuo) $P_j^{(i)}$ referido a *p* ejes, el escalamiento, rotación y traslación pueden expresarse algebraicamente por la transformación;

$$X_i \rightarrow \rho_i X_i H_i + T_i$$

en la cual la matriz ortogonal de rotación H_i, el factor de escala p_i y la matriz de traslación T_i, se hallarán de forma que se minimice,

$$S_r = \sum_{j=1}^{n}\sum_{i=1}^{m} \Delta^2 \left(P_j^{(i)}, G_i \right)$$

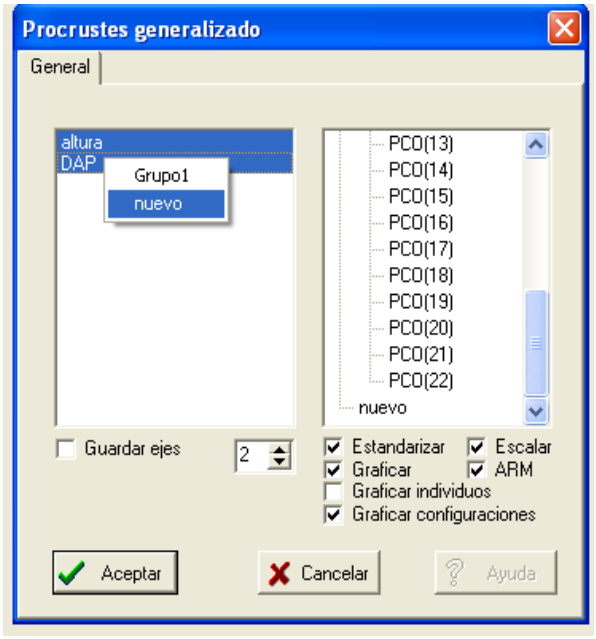
donde Δ(A,B) es la distancia euclídea entre el par de puntos A y B, y G_j es el centroide de los m puntos análogos $P_j^{(i)}$ (i=1, 2,...,*m*)

En síntesis la técnica de Procrustes generalizado provee un método para consensuar ordenaciones que involucra tres acciones:
1. Traslación (centrado de las ordenaciones)
2. Rotación (de las ordenaciones para minimizar la diferencia entre ellas)
3. Escalamiento (multiplicación por factores para minimizar diferencias de tamaño).

Ejemplo 41: El archivo procrustes.idb contiene un conjunto de datos de 23 individuos sobre los que se registraron dos tipos de información, genética y fenotípica. Los datos genéticos provienen de marcadores moleculares de ADN y son del tipo binario, mientras que los datos fenotípicos son variables continuas. Se trabajó sobre un conjunto de 239 marcadores moleculares (datos genéticos) y dos variables fenotípicos (altura de planta y diámetro a la altura del pecho o DAP). Mediante el Análisis de Procrustes se busca cuantificar el consenso entre la ordenación de los individuos obtenida por Análisis de Coordenadas Principales de la matriz de distancias genéticas y la ordenación de los mismos individuos mediante los de datos fenotípicos. En una primera instancia se solicitó en Análisis de Coordenadas principales usando los datos de las 239 variables genéticas y se guardaron los 22 ejes resultantes.

Menú ESTADÍSTICA ⇒ ANÁLISIS MULTIVARIADO ⇒ PROCRUSTES GENERALIZADO, provee una ventana donde se seleccionan las variables del análisis, en este caso las 22 coordenadas principales y las dos variables morfológicas (sobre este tipo de variables no se realizo una técnica de reducción de dimensión ya que eran solo dos). Posterirormente, deben agruparse las variables en función del tipo de información que brindan, en este ejemplo, las coordenadas principales (que provienen de información molecular) se asignaron a un grupo (Grupo1) y las variables que contienen la información morfológica a otro grupo (Grupo2). Para realizar esta asignación, deben seleccionarse las variables y con el botón derecho del ratón indicar nuevo, automáticamente InfoStat colocará

en la ventana de la derecha todas las variables seleccionadas en un mismo grupo, luego se vuelve a realizar la operación para asignar las variables restantes a otro grupo. En la segunda instancia la ventana será la siguiente:



Si se selecciona la opción ARM, InfoStat graficará un árbol de recorrido mínimo sobre la configuración de consenso.

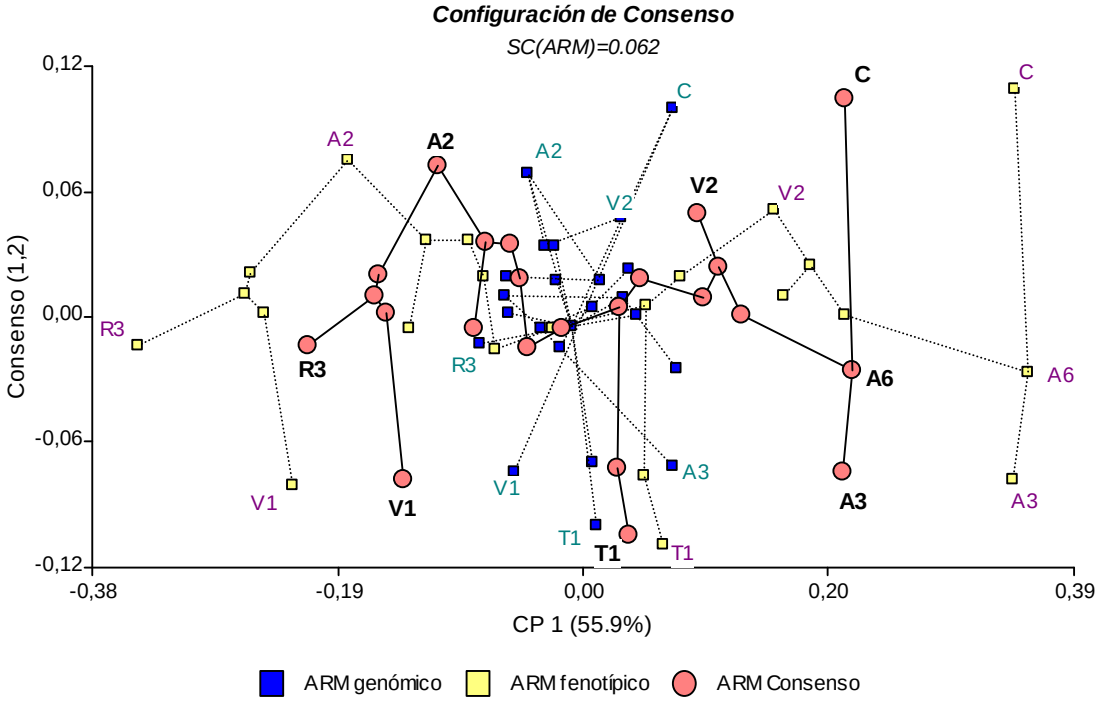


Figura 34: ARM de configuraciones originales y de consenso. Archivo Procrustes.idb.

Tabla 64: Procrustes generalizado. Archivo Procrustes.idb.

Cuadro de Análisis de la Varianza

Sumas de cuadrados dentro por caso

	Consenso	Residual	Total
A1	0.026	0.023	0.049
A2	0.052	0.029	0.081
A3	0.114	0.052	0.166
A4	0.065	0.039	0.104
A5	0.069	0.040	0.109
A6	0.113	0.057	0.169
B1	0.041	0.029	0.070
B2	0.032	0.020	0.052
C	0.122	0.050	0.173
Z	0.043	0.027	0.070
N1	0.034	0.025	0.059
N2	0.025	0.023	0.048
R1	0.046	0.031	0.077
R2	0.024	0.023	0.046
R3	0.022	0.022	0.044
R4	0.108	0.055	0.163
R5	0.071	0.041	0.111
R6	0.034	0.027	0.061
R7	0.027	0.023	0.050
T1	0.053	0.034	0.088
T2	0.041	0.018	0.060
V1	0.029	0.023	0.052
V2	0.067	0.033	0.100
Total	1.257	0.743	2.000

Sumas de cuadrados dentro por grupo

	Consenso	Residual	Total
Grupo1	0.628	0.372	1.000
Grupo2	0.628	0.372	1.000
Total	1.257	0.743	2.000

El consenso entre la ordenación producida por matriz de datos genéticos y la obtenida a partir de datos morfológicos es del 63%, este se obtiene dividiendo el total (2) por el consenso total obtenido (2/1.257=0.6285).

Series de Tiempo

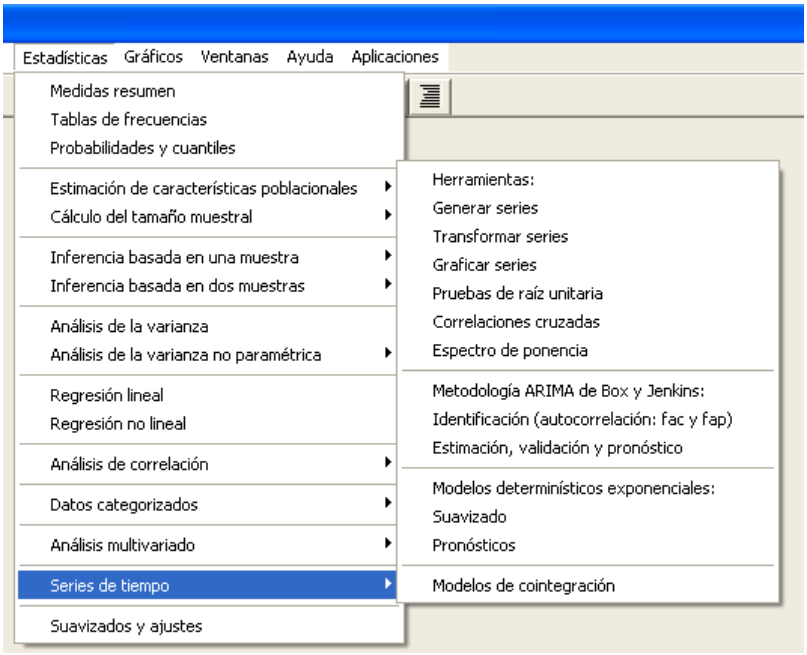
El módulo Series de Tiempo de InfoStat aborda el análisis de datos observados en forma secuencial a intervalos regulares de tiempo. Se han implementado dos enfoques clásicos para el modelado de una serie y el pronóstico de valores futuros de la serie: 1) técnicas determinísticas de suavizado (ver Estadísticas, Suavizados y ajustes) y 2) técnicas basadas en los modelos de series de tiempo ARIMA de Box y Jenkins (1976).

En esta versión, InfoStat permite el análisis de series de tiempo univariadas (la realización de un proceso estocástico definido en los números reales). Cuando en la tabla de datos la serie se ha ingresado como una columna, InfoStat interpreta que la secuencia en el tiempo de los datos viene dada por el orden en que han sido ingresados (número de casos). El usuario puede, alternativamente, utilizar una columna adicional para indexar en el tiempo las observaciones de la columna que contiene la serie (fechado).

El modelado propiamente dicho de una serie de tiempo, en el caso de las técnicas determinísticas de suavizado, depende fuertemente de la elección, basada en el juicio crítico del usuario, del o de los *parámetros* que gobiernan el suavizado.

InfoStat admite series con datos faltantes. El usuario puede solicitar la predicción automatizada de los datos faltantes.

El tamaño del archivo de datos, es decir el número de casos y el número de series para trabajar en este módulo, sólo depende de la capacidad de memoria RAM de la computadora personal en la que se ejecuta InfoStat.



InfoStat ofrece en este módulo la posibilidad de construir los gráficos comúnmente utilizados para representar series de tiempo, sin necesidad de recurrir al menú Gráficos.

Además, el usuario dispone facilidades para simular una amplia gama de procesos generadores de series de tiempo, incluyendo procesos estocásticos *estacionarios* y *no*

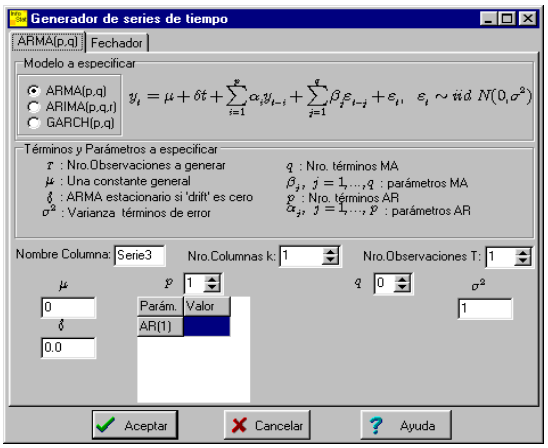
estacionarios, *estacionales* y *no estacionales*. Las series simuladas son automáticamente guardadas como columnas de la tabla de datos activa.

En el caso de los modelos ARIMA de Box y Jenkins, InfoStat ofrece al usuario trabajar siguiendo los tres pasos fundamentales de *identificación*, *estimación* y *validación*. Para cualquiera de estas estrategias, los pronósticos de interés se pueden construir de forma ágil y con capacidades gráficas de representación altamente informativas. InfoStat posibilita estimar la función de impulso-respuesta y sus bandas de confianza para describir el comportamiento del proceso frente a un disturbio instantáneo.

Menú ESTADISTICAS ⇒ SERIES DE TIEMPO permite acceder a los submenúes que se muestran a continuación:

Simulación y transformaciones

En el bloque **Herramientas** de submenúes, InfoStat permite generar series de tiempo por simulación, transformar series, graficarlas, realizar la prueba de raíces unitarias y obtener funciones de correlación cruzada entre dos series de tiempo.



Al invocar el submenú **Generar series** se activa la ventana **Generador de series de tiempo** con dos solapas, **ARMA(p,q)** y **Fechador**. En la solapa **ARMA(p,q)** se deben especificar los parámetros del modelo que simula el proceso generatriz de una serie de tiempo de *T* observaciones. En la solapa **Fechador** se puede indicar la estructura que indexa la serie de tiempo generada (frecuencia, períodos, inicio y fin del espacio de tiempo de la serie generada).

Los modelos que se pueden simular son los modelos estacionarios ARMA(p,q), no estacionarios ARIMA(p,q,d) (ver menú Metodología ARIMA de Box Jenkins), y los heteroscedásticos condicionales GARCH(p,q) (ver Hamilton, 1994). Cuando el usuario selecciona el tipo de modelo que desea usar se presenta automáticamente la ecuación del modelo correspondiente y la explicación de los términos y parámetros a especificar. En la figura anterior se puede observar la ecuación del modelo ARMA(p,q) y los términos y parámetros a especificar para el mismo.

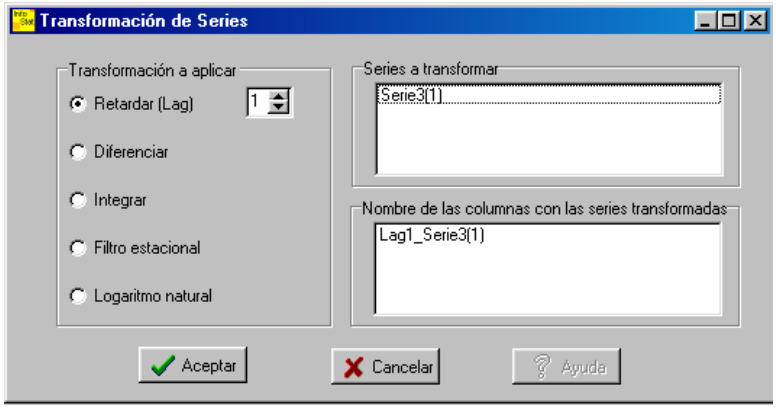
En el campo **Nombre columna** InfoStat sugerirá un nombre para la serie a generar que el usuario podrá cambiar. En la figura anterior InfoStat está sugiriendo nombrar a la serie como “Serie3” dado que la tabla de datos activa ya tenía en uso las dos primeras columnas.

En el campo **Nro Columnas k** debe especificarse cuantas series de tiempo se desean generar. En el campo **Nro. Observaciones T** debe ingresarse la longitud de la/s serie/s a generar.

Los botones **p** y **q** posibilitan especificar los *órdenes de las porciones autorregresivas (AR)* y de *medias móviles (MA),* respectivamente. Asociado a cada botón existe una grilla para ingresar los parámetros correspondientes. En la figura anterior, como **p** es igual 1 y **q** es igual a cero, sólo se visualiza la grilla de parámetros que deberán especificarse para la porción AR. A medida que **p** y **q** son incrementados, las grillas habilitarán tantas filas como parámetros debieran ser especificados por el usuario. Los campos μ, δ y σ^2 deberán ser completados con los valores de los parámetros "constante", "tendencia" y "varianza" del proceso generatriz.

Para los modelos ARIMA(p,q,d) y GARCH(p,q) el usuario deberá proceder de forma análoga a la descripta para el modelo ARMA(p,q).

Al invocar el submenú **Transformar series** se presentará en pantalla el Selector de Variables donde el usuario deberá especificar cuál es o son las columnas del archivo (series) que se desean transformar. Al seleccionar una serie se presentará la ventana **Transformación de Series**, en la que se muestra la serie seleccionada para aplicar la transformación y el nombre de la columna que InfoStat adjudicará a la serie transformada. Este nombre puede ser cambiado por el usuario posicionándose sobre el nombre que desea modificar.



Las transformaciones posibles de ser utilizadas incluyen: retardos con *lags* especificados por el usuario, diferencias, integraciones, filtros estacionales y transformación logaritmo (Pindyck y Rubinfeld, 1999). Para realizar dos o más diferenciaciones, el usuario deberá repetir la transformación cuantas veces sea necesaria. En el caso de aplicar un **Filtro estacional** se habilita un campo para indicar la amplitud del ciclo (mensual,1; bimensual, 2; ...; anual, 12).

Al invocar el submenú **Graficar series** se presentará en pantalla el Selector de Variables donde el usuario deberá especificar cuál es o son las columnas del archivo que se desean graficar; opcionalmente el usuario puede seleccionar una variable representativa del tiempo para asociar al eje de las X. Al **Aceptar**, InfoStat construye automáticamente el diagrama de dispersión de la serie de tiempo y presenta la ventana de herramientas gráficas a partir de la cual el usuario puede adecuar el gráfico a sus necesidades. Esta ventana permite aplicar las funciones gráficas usuales del menú Gráficos de InfoStat con pequeñas excepciones como son aquellas que permiten especificar el número de puntos a visualizar en el eje X (en series

de tiempo generalmente existe un número alto de valores para X, InfoStat permite decidir los valores de la escala que se harán visibles a través de la especificación del número de observaciones que quedarán comprendidas entre dos *ticks*).

En la Figura 35 se muestra la gráfica obtenida por defecto para una serie generada por simulación a partir de un modelo AR(1) con constante cero, sin tendencia y varianza uno.

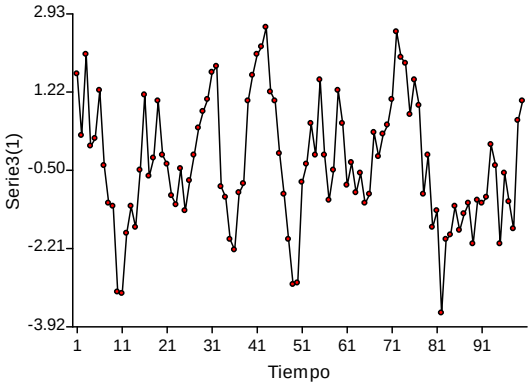


Figura 35: Serie para un modelo AR(1), generada por simulación.

Prueba de raíz unitaria

Al invocar el submenú **Prueba de raíz unitaria** se presentará en pantalla el Selector de Variables donde el usuario deberá especificar cuál es la columna del archivo (serie) para la que se desea realizar la prueba. Al **Aceptar**, InfoStat presentará el valor de varios estadísticos para la prueba y el valor de probabilidad asociado a cada uno de ellos. La hipótesis nula postula que la serie tiene una raíz unitaria, la alternativa que establece la ausencia de raíz unitaria, es decir la estacionaridad del proceso.

InfoStat implementa tres pruebas estadísticas para la hipótesis de raíz unitaria. Estas son la prueba de Dickey-Fuller (1979), Dickey-Fuller aumentada (1981) y la de Phillips-Perron (1988) (ver Hamilton, 1994).

Dickey y Fuller observaron que bajo la hipótesis alternativa, cuando existe una única raíz unitaria, la serie diferencia Δy_t es estacionaria. Simbólicamente,

$$y_t = \alpha y_{t-1} + u_t, \quad |\alpha| < 1$$

$$y_t - y_{t-1} = (\alpha - 1) y_{t-1} + u_t,$$

$$\Delta y_t = \gamma y_{t-1} + u_t, \quad \gamma < 0.$$

Luego, la hipótesis de raíz unitaria puede reescribirse como:

$$H_0 : \gamma = 0$$

$$H_1 : \gamma < 0$$

Para realizar la prueba de Dickey y Fuller, InfoStat estima γ mediante mínimos cuadrados ordinarios y genera los valores de probabilidad asociados a partir de la distribución empírica del estadístico obtenida por simulación Monte Carlo. La distribución dependerá de la existencia de una constante, de una tendencia o de ambas en el modelo, por lo que InfoStat presentará las pruebas bajo todas estas situaciones en forma automática. Es importante observar que si $y_t \sim AR(p)$ con una raíz unitaria luego la serie diferenciada $\Delta y_t \sim AR(p-1)$. Por ello, la prueba de Dickey y Fuller aumentada se realiza en base a la estimación por mínimos cuadrados ordinarios del siguiente modelo:

$$\Delta y_t = \gamma y_{t-1} + \sum_{j=1}^{p-1} \beta_j \Delta y_{t-j} + u_t$$

Al igual que para la prueba básica de Dickey y Fuller, los valores *p* reportados en este caso son obtenidos de las distribuciones empíricas del estadístico derivadas por simulación Monte Carlo, las que dependen si el modelo aumentado tiene una constante, una tendencia, o ambas. En la prueba propuesta por Phillips y Perron se corrigen las pruebas de Dickey y Fuller para situaciones en que los términos de error están serialmente correlacionados y/o son heterocedásticos.

Correlaciones cruzadas

El submenú **Correlaciones Cruzadas** muestra el ***Selector de Series*** donde el usuario deberá especificar cuáles son las dos series para las cuales se desea obtener la función de correlación cruzada. La función de correlación cruzada presenta la correlación entre ambas series para distintos *lags* de la segunda serie respecto de la primera. Conceptualmente la función de correlación cruzada es análoga a la función de autocorrelación, excepto que las correlaciones no son obtenidas entre observaciones de una misma serie sino entre dos series diferentes. Permite obtener la función de correlación cruzada entre dos series de tiempo. La **correlación cruzada para el lag k** mide la magnitud de la correlación linear entre los valores de la primera serie y los valores de la segunda serie, k periodos hacia adelante. InfoStat considera como primera serie a la columna de la tabla activa que ha sido seleccionada primero en la ventana del selector de variables. La función de correlación cruzada resultante se muestra automáticamente en la ventana de resultados y en la grafica. En la ventana **Resultados** se presenta, además de la correlación para cada lag, su error estándar, el estadístico T y el valor p de la prueba de hipótesis de correlación cero para ese lag. En la ventana **Grafica** se representan también las bandas de confianza al 95% para la función de correlación cruzada. El usuario puede modificar el nivel de confianza para estos intervalos. La función de correlación cruzada entre dos series suele ser usada para determinar si la segunda serie podría ayudar a predecir la primera.

La función de covarianza cruzada entre dos series de tiempo $x_{1,t}$ y $x_{2,t}$, $t = 1,...,T$, para el retardo *k*, denotada como $C_{12}(k)$, es estimada en InfoStat de la siguiente forma,

$$C_{12}(k) = \frac{1}{T-k}\sum_{t=1+k}^{N-k} x_{1,t}x_{2,t-k}, \quad k = -([N/4]+1),..,-1,0,1,...,[N/2]+1,$$

donde [.] representa la parte entera del argumento. Al dividir $C_{12}(k)$ por la raíz cuadrada del producto de las funciones de autocovarianza de cada serie en el retardo cero, $(C_{11}(0)C_{22}(0))^{1/2}$, se obtiene el coeficiente de la función de correlación cruzada para los datos muestrales,

$$\rho_{12}(k) = \frac{C_{12}(k)}{(C_{11}(0)C_{22}(0))^{1/2}}$$

Usando el archivo de datos *CrossCorr.idb*, InfoStat genera la siguiente salida y gráfico de la función de correlación cruzada.

Tabla 65: Función de correlación cruzada entre las series x e y

Función de correlación cruzada

Información general

Serie	Nro.Obs.	Media	Var(n-1)	D.E.
x	10	14.80	6.84	2.62
y	10	13.70	4.23	2.06

Función de correlación cruzada: r(Lag)

Lag	Coef	E.E.	T	p	Signif
-3	0.08	0.52	0.16	0.8728	
-2	0.36	0.49	0.73	0.4687	
-1	0.56	0.44	1.25	0.2145	
0	0.92	0.32	2.90	0.0052	*
1	0.69	0.32	2.18	0.0331	*
2	0.47	0.36	1.31	0.1955	
3	0.36	0.43	0.82	0.4156	

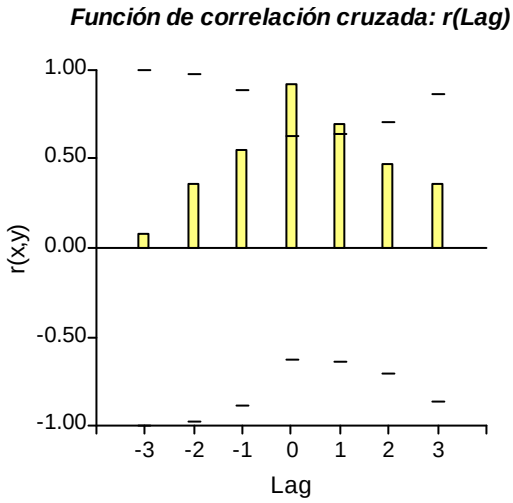


Figura 36: Representación de una realización de una caminata aleatoria (proceso no estacionario) y de su diferenciación (proceso estacionario).

Espectro de potencia

Al invocar el submenú **Espectro de potencia** se presentará en pantalla el Selector de Variables donde el usuario deberá seleccionar las series para las cuales se desea obtener el espectro de potencia. El análisis espectral es utilizado en distintas disciplinas para particionar la varianza de una serie de tiempo en función de las frecuencias. Para series de tiempo estocásticas las contribuciones de las diferentes frecuencias a la varianza son medidas en términos de la densidad espectral o espectro de potencia.

La palabra "espectro" proviene de la óptica. Los colores rojo, banco y azul del espectro electromagnético son utilizados a menudo para describir la distribución de frecuencias del espectro. Un espectro cuya densidad espectral decrece con frecuencias crecientes es llamado "un espectro rojo", por analogía con la luz visible, en la que el rojo corresponde a longitudes de ondas largas (frecuencias bajas). De forma similar, un espectro cuya magnitud crece con la frecuencia es llamado "un espectro azul". Un "espectro blanco" es uno en el que las componentes espectrales tienen aproximadamente las misma amplitudes a lo lardo del rango de frecuencias. Así, series de tiempo que presentan periodos largos de variabilidad tienden a presentar espectros rojos, como en el caso de los ciclos económicos de largo plazo., en tanto que espectros blancos suelen presentarse en los errores de medición de instrumentos de laboratorio.

InfoStat estima el espectro de potencia utilizando un método espectral no paramétrico, basado en la transformación de Fourier de la función de autocovarianza de una series de tiempo y_t, $t = 1,...,T$, que viene dada por:

$$G_k = \sum_{m=0}^{M} C_{yy}(m) e^{-i2\pi km/M}, \; k = 0,...,T/2,$$

donde $C_{yy}(m) = \frac{1}{T}\sum_{t=1}^{T-m} y_t y_{t-m}$ es un estimador (sesgado) de la función de autocovarianza, estimada para u total de *M* retardos. Se utiliza este estimador sesgado por cuanto conside con la densidad espectral estimada por la transformada rápida de Fourier (FFT) de la serie original (Emery y Thompson, 1997). Como $C_{yy}(m)$ es una función par, el espectro de $\{y_t\}$ es estimado en InfoStat por medio de la transfomación coseno,

$$G_k = C_{yy}(0) + 2\sum_{m=1}^{M} C_{yy}(m)\cos\frac{2\pi km}{T},\ k = 0,...,T/2,$$

donde G_k está centrada en frecuencias positivas $f_k = k/T$ y el intervalo de Nyquist $0 \le f_k \le f_T = 1/2$ es dividido en T/2 segmentos (*T* es par).

El archivo *especpote.idb* contiene datos de las temperaturas medias mensuales (grados centígrados) de la superficie del mar en un cierto punto de coordenadas 48°55.16'N, 125°32.17', tomados desde Enero de 1982 hasa Diciembre de 1984. El espectro de potencia estimado utilizando InfoStat se reproduce en la siguiente tabla:

Tabla 66: Análisis del espectro de potencia de la serie SST

Espectro de potencia

Información general

Serie	Nro.Obs.	Media	Var(n-1)	D.E.
SST	36	10.811	4.274	2.067

Espectro de potencia

Frecuencia	Coef
0.000	0.000
0.028	5.292
0.056	1.866
0.083	62.934
0.111	0.455
0.139	0.002
0.167	2.750
0.194	0.030
0.222	0.151
0.250	0.103
0.278	0.072
0.306	0.103
0.333	0.267
0.361	0.302
0.389	0.031
0.417	0.016
0.444	0.045
0.472	0.177
0.500	0.401

El gráfico del espectro de potencia de este problema es:

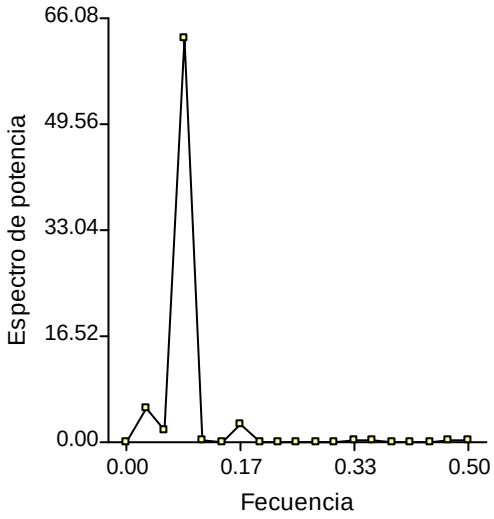


Figura 37: Representación del espectro de potencia de la serie SST

De la tabla y el gráfico puede observarse un pico en el espectro centrado aproximadamente en la frecuencia anual, por cuanto en la escala mensual del gráfico corresponde a una frecuencia de 0.083 ciclos por mes (0.083x12=1 ciclo por año).

Metodología ARIMA de Box y Jenkins

InfoStat permite aplicar la metodología propuesta por Box y Jenkins (1976) para identificar, estimar y validar modelos auto-regresivos integrados de promedio móviles (ARIMA). Un modelo ARIMA representa una expresión algebraica que establece cómo las observaciones sobre una variable, en un determinado momento de tiempo, se relacionan estadísticamente con observaciones registradas en el pasado sobre la misma variable. La construcción de modelos ARIMA requiere trabajar con suficiente información. Como en todo proceso de modelización un buen modelo será aquel más parsimonioso (menor orden) que muestra un buen ajuste a los datos.

El proceso y_t tiene representación ARIMA(p.q,d) si la diferencia de orden d del mismo, $\Delta^d y_t$, tiene representación ARMA (p,q). Diremos que y_t tiene una representación ARMA(p,q) si $y_t = \frac{\beta(L)}{\alpha(L)}\varepsilon_t$, donde $\alpha(L)$ es un polinomio AR de orden q, $\beta(L)$ es un polinomio MA de orden p y ε_t es el término de error en el momento t-ésimo (Box y Jenkins, 1976). Si q=0 luego diremos que $y_t \sim AR(p)$ y si p=0 que $y_t \sim MA(q)$.

Por ejemplo, el proceso estocástico $\{y_t\}$ tendrá una representación AR(1), MA(1) ó ARMA(1,1) si puede ser escrito como:

(a) $y_t = \alpha y_{t-1} + \varepsilon_t$,

(b) $y_t = \varepsilon_t + \beta\varepsilon_{t-1}$,

(c) $y_t = \alpha y_{t-1} + \varepsilon_t + \beta\varepsilon_{t-1}$,

respectivamente, donde la secuencia $\{\varepsilon_t\}$ forma una martingala en diferencia (MDS) condicionalmente homoscedástica con varianza incondicional σ^2 (Hamilton, 1994).

Una serie estacionaria básicamente es aquella que tiene media, varianza y función de autocorrelación constantes en el tiempo. Series de mayor longitud debieran ser usadas cuando no se puede sostener el supuesto de estacionaridad del proceso. Series no estacionarias pueden ser transformadas, mediante funciones simples, en series estacionarias. InfoStat permite realizar la transformación *diferencia de series* previamente a la construcción del modelo.

La **Metodología de Box-Jenkins** se basa en los siguientes pasos:

En base a las funciones de autocorrelación y de autocorrelación parcial elegir *p* y *q* (paso conocido como identificación del modelo),

Estimar los parámetros (usando mínimos cuadrados ordinarios, máxima verosimilitud, etc.)

Realizar un control diagnóstico sobre los residuos del modelo (buscando falta de correlación serial, normalidad, homogeneidad de varianzas y estacionariedad).

Si el diagnóstico sugiere un buen ajuste del modelo propuesto, se podrá realizar pronósticos. En caso contrario, se deberá repetir la estimación y diagnóstico con diferentes valores de *p* y *q*.

Como herramientas para la identificación de modelos ARMA, InfoStat provee la posibilidad de obtener:

Gráficos de la serie observada vs. el tiempo. A partir de estos, el usuario debiera responder si es necesario diferenciar y/o remover tendencias. Si se diferencia, el modelo será estimado sobre la serie diferencia. La serie diferenciada debiera ser estacionaria.

Ejemplo 42: En el siguiente gráfico se representa una serie de 40 observaciones de un proceso no estacionario (caminata aleatoria) y la serie diferenciada (caminata aleatoria diferenciada). La serie no estacionaria fue simulada con InfoStat usando el siguiente modelo: $y_t = y_{t-1} + u_t,\ u_t = 0.3u_{t-1} + 0.5\varepsilon_t + \varepsilon_t,\ \varepsilon_t \sim iid\ N(0,\sigma^2)$. *La diferenciación produce una serie estacionaria.*

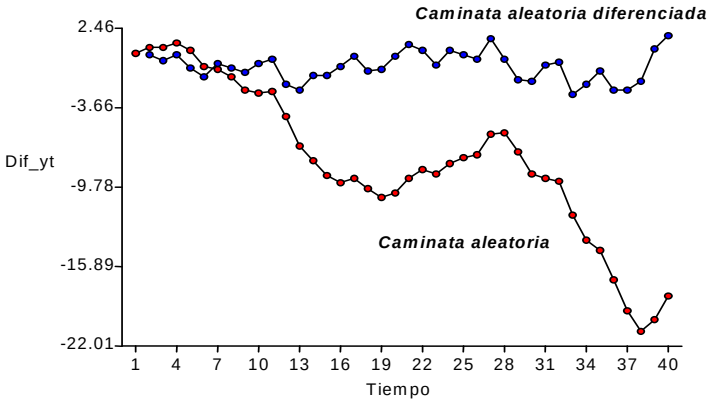


Figura 38: Representación de una realización de una caminata aleatoria (proceso no estacionario) y de su diferenciación (proceso estacionario).

Ejemplo 43: Si el proceso que genera los datos presenta una tendencia se podría proponer el siguiente modelo:

$$y_t = \lambda_0 + \lambda_1 t + \frac{\beta(L)}{\alpha(L)} u_t,$$

En el gráfico siguiente se representa una serie generada en InfoStat a partir del modelo $y_t = 5 + 1.5t + u_t,\ u_t = 0.3u_{t-1} + 0.5\varepsilon_{t-1} + \varepsilon_t,\ \varepsilon_t \sim iid\ N(0,3)$ y la tendencia pura determinística:

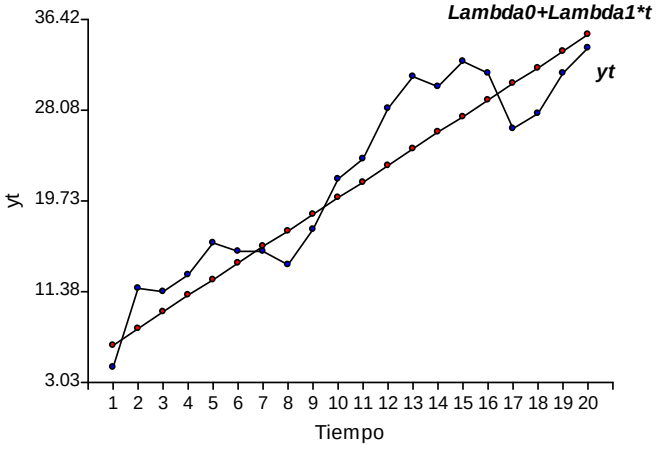


Figura 39: Serie generada a partir del modelo del Ejemplo 42

En casos como el de este ejemplo, se debería eliminar la "tendencia" mediante la estimación en primera instancia de los parámetros de la regresión $y_t = \lambda_0 + \lambda_1 t + \varepsilon_t$ y luego asumir que:

$$y_t - \hat{\lambda}_0 - \hat{\lambda}_1 t \sim ARMA(p,q)$$

Representaciones gráficas de las funciones de autocorrelación y de autocorrelación parcial: El submenú Identificación (autocorrelación: fac y facp), permite obtener automáticamente, para cada serie univariada cuyo modelo debe ser identificado, las funciones de autocorrelación (fac) y autocorrelación parcial (facp).

La **función de autocorrelación** es una forma de medir cómo se correlacionan las observaciones dentro de una misma serie de tiempo. A partir de estas funciones, el usuario puede resumir las relaciones estadísticas significativas a través de la selección de uno de los modelos de la familia ARIMA ya que cada modelo ARIMA tiene un par de funciones de autocorrelación y autocorrelación parcial asociadas.

Si γ_j es la j-ésima autocovarianza y γ_0 es la varianza luego $\rho_j = \frac{\gamma_j}{\gamma_0}$ es la j-ésima autocorrelación que se puede estimar como sigue:

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0},\ \hat{\gamma}_j = \frac{1}{T-j}\sum_{t=j+1}^{T}(y_t - \overline{y})(y_{t-j} - \overline{y})\,.$$

El gráfico de los coeficientes de autocorrelación para distintos *lags* provee una representación de la función de autocorrelación. El comportamiento de la función de autocorrelación puede describirse de la siguiente manera:

Si el proceso es un AR puro, las autocorrelaciones declinarán exponencialmente;

Si el proceso es un MA puro declinará a cero rápidamente, dependiendo del orden q del proceso;

Si el proceso es un ARMA(p,q), declinará rápidamente.

Ejemplo 44: Se presenta la fac de un proceso simulado AR (1). La serie fue generada usando el generador estocástico de InfoStat con un coeficiente auto-regresivo AR(1) igual a 0.75. Las barras representan los coeficientes de autocorrelación muestral y los guiones inferiores y superiores a la línea de referencia los límites del intervalo 95% de confianza para el coeficiente de autocorrelación poblacional.

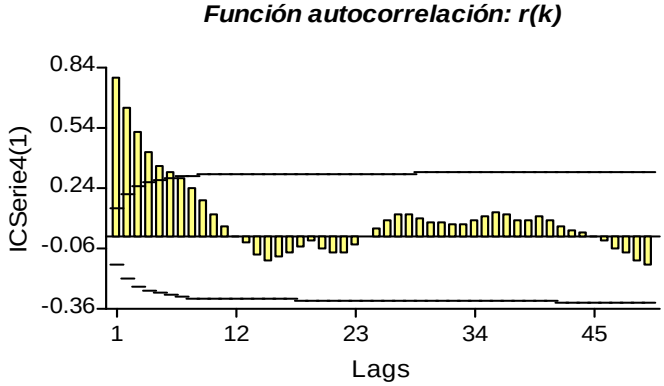


Figura 40: Función de autocorrelación.

Para conceptualizar la **función de autocorrelación parcial** (fap) supongamos que y_t es de covarianza estacionaria, y que se ajusta la regresión de y_t sobre y_{t-1}, esto es $y_t = \phi_{11} y_{t-1} + \varepsilon_t$. Luego ϕ_{11} es la primera autocorrelación parcial (la que puede ser estimada por mínimos cuadrados ordinarios).

Cuando se realice la regresión de y_t sobre y_{t-1} e y_{t-2} se obtendrá la segunda autocorrelación parcial ϕ_{22}: $y_t = \phi_{21} y_{t-1} + \phi_{22} y_{t-2} + \varepsilon_t$, y así sucesivamente.

El gráfico de los coeficientes de autocorrelación parcial para distintos *lags* provee una representación de la función de autocorrelación parcial. El comportamiento de la función de autocorrelación parcial puede describirse de la siguiente manera:

Si el proceso es un AR puro, las autocorrelaciones parciales declinarán rápidamente, dependiendo del oden *p* del proceso;

Si el proceso es un MA puro las autocorrelaciones parciales declinarán a cero suavemente

Si el proceso es un ARMA(p,q), declinará rápidamente.

Ejemplo 45: Se presenta la fap de un proceso AR (1) con coeficiente de autoregresivo igual a 0.75, generado usando el simulador de InfoStat. Las barras representan los coeficientes de autocorrelación parcial muestral y las líneas inferiores y superiores que demarcan los límites del intervalo 95% de confianza para el coeficiente de autocorrelación parcial poblacional.

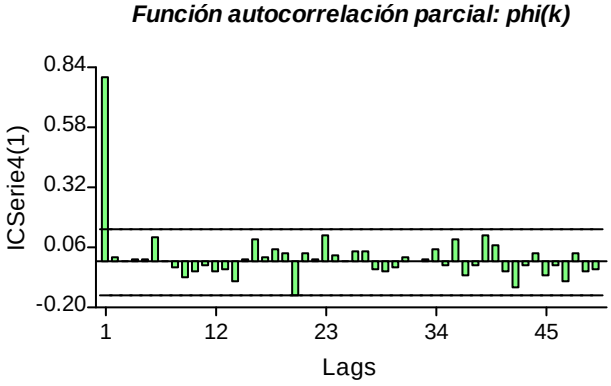


Figura 41: Función de autocorrelación parcial.

En la práctica estimar *T/4* autocorrelaciones y autocorrelaciones parciales mediante mínimos cuadrados ordinarios puede ser computacionalmente costoso. InfoStat realiza dicha estimación mediante el algoritmo recursivo de Durbin (1960), basado en resolver el sistema de ecuaciones de Yule-Walker:

$$\hat{\phi}_{11} = \hat{\rho}_1$$

$$\hat{\phi}_{kk} = \frac{\hat{\rho}_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j}\hat{\rho}_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j}\hat{\rho}_j} \quad (k = 2,3,...)$$

donde $\hat{\phi}_{k,j} = \hat{\phi}_{k-1,j} - \hat{\phi}_{k,k}\hat{\phi}_{k-1,k-j} \quad (k = 3,4,...; j = 1,2,...,k-1)$.

Criterios de Información: Además de las herramientas del proceso de identificación mencionadas anteriormente, InfoStat calcula de forma automática del proceso de estimación criterios estadísticos de información (SIC), que pueden ser utilizados tanto para identificación como para diagnóstico.

La idea subyacente de estos criterios es el hallar *p* y *q* de forma tal de minimizar la suma de cuadrados residual sujeta a términos de penalización que dependen de *p* y *q* para reducir el potencial de incurrir en sobreajustes o ajuste de modelos poco parsimoniosos. Los criterios disponibles difieren en la forma en que se penaliza la sobreparametrización.

La metodología de trabajo basada en los SIC como herramienta de identificación (para elegir *p* y *q*) sugiere:

(1) Elegir $\tilde{P}$ y $\tilde{Q}$ los máximos posibles para *p* y *q* (información que puede obtenerse de las *fac* y *facp* en forma conjunta). En la práctica es conveniente tomar $p+q<4$.

(2) Para cada modelo $ARMA(p,q)$ tal que $p \leq \tilde{P}$ y $q \leq \tilde{Q}$ estimar los parámetros y obtener los residuos $\hat{u}_t(p,q)$.

(3) Calcular $\hat{\sigma}^2(p,q) = T^{-1}\sum_{t=1}^{T}\hat{u}_t^2(p,q)$.

(4) Calcular el SIC elegido.

(5) Elegir *p* y *q* tal que minimice el SIC elegido.

Los criterios de información implementados en InfoStat son:

Akaike: $AIC = \ln\sigma^2(p,q) + \frac{2(p+q)}{T}$.

Schwartz: $BIC = \ln\sigma^2(p,q) + \frac{(p+q)\ln T}{T}$.

Hannan-Quinn: $HQ = \ln\sigma^2(p,q) + \frac{c(p+q)\ln(\ln T)}{T}, c > 1.$

Nota:

(1) Para *T* grandes, AIC elegirá los *p* y *q* "correctos", pero puede sobreparametrizar para *T* no tan grande, si AR(2) es el modelo, por ejemplo, no elegirá *p*=0 o *p*=1, pero si $p \geq 2$.

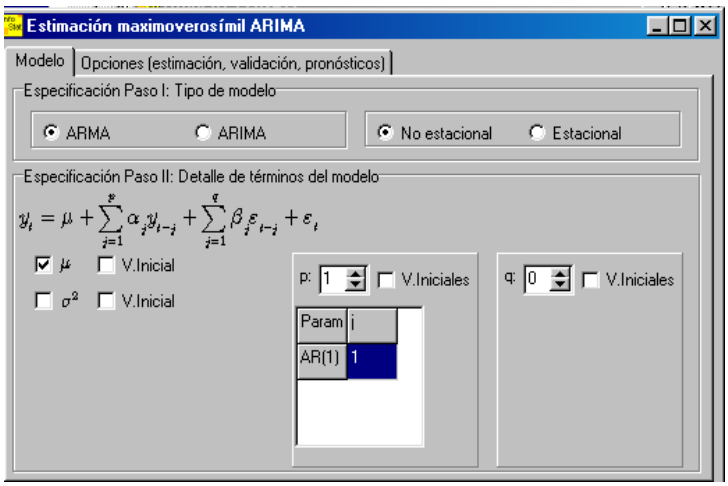
(2) En muestras finitas BIC tiende a subestimar *p* y *q*, ya que el término de penalidad tiende a dominar.

(3) HQ soluciona (o trata de solucionar) el problema de BIC en muestras finitas.

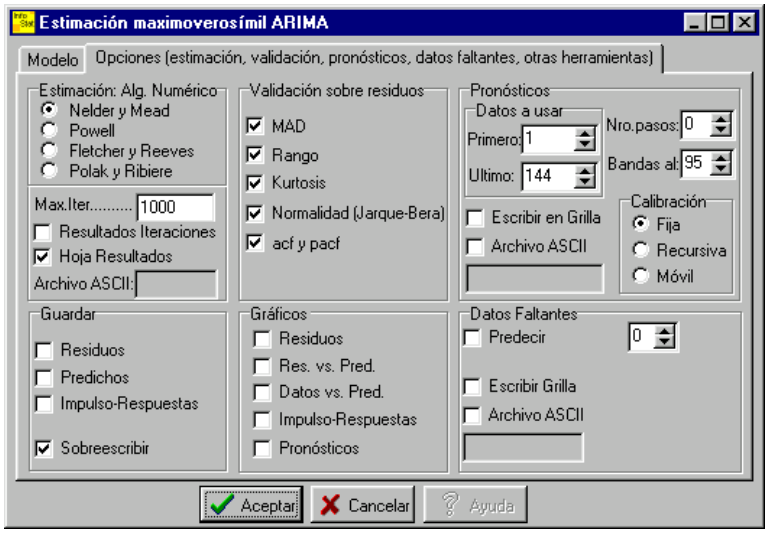
El submenú **Estimación, validación y pronósticos**, implementa el segundo y tercer paso de la metodología de Box y Jenkins una vez que se ha identificado el modelo. Al invocar este submenú aparecerá una ventana con dos solapas: **Modelo** y **Estimación, validación, pronósticos, pronóstico, datos faltantes, otras herramientas.** La solapa **Modelo** permite especificar la ecuación del modelo a estimar. Se puede seleccionar modelos estacionarios ARMA y modelos no estacionarios ARIMA (en tal caso InfoStat requerirá que se ingrese el valor de *d*, el parámetro de diferenciación a utilizar). Los modelos a estimar pueden ser **No estacionales** (sin tendencia) o **estacionales** (en tal caso InfoStat solicitará se ingrese la frecuencia que describe la estacionalidad; InfoStat también admite la presencia de ciclos menores dentro de un ciclo mayor de estacionalidad).

De acuerdo con las especificaciones realizadas por el usuario se mostrará la ecuación del modelo a estimar y se podrán leer en la misma pantalla el detalle de los términos del modelo. Además, se activarán tantos campos como parámetros se deban estimar para que el usuario ingrese los valores iniciales de los mismos a considerar en un proceso de estimación iterativo. Se sugiere en un primer paso activar el campo relacionado al parámetro μ

(esperanza de la serie), en tal caso InfoStat asume como valor de partida para dicho parámetro en el proceso de estimación el valor redondeado de la media aritmética de la serie.



InfoStat realiza la estimación de los parámetros del modelo mediante máxima verosimilitud condicionada a los *p* primeros valores observados de la serie (Hamilton, 1994). La solapa **Estimación, validación, pronósticos, datos faltantes, otras herramientas** permite realizar controles sobre los procedimientos de estimación, indicar cuales serán las herramientas utilizadas para el diagnóstico del modelo estimado y realizar pronósticos cuando el modelo ya ha sido validado.



La función de verosimilitud es optimizada numéricamente con algunos de los siguientes algoritmos (**Estimación Alg. Numérico**), según indique el usuario: **Nelder y Mead**, **Powell**, **Fletcher y Reeves** y **Polak y Ribiere**. Los dos primeros se basan en el procedimiento *Downhill* y los dos últimos se basan en el método del gradiente (Press *et al*., 1986). Se recomienda en una primera instancia seleccionar el algoritmo de Nelder y Mead con los valores iniciales de μ sugeridos por InfoStat. Posteriormente se podría aplicar el

algoritmo de Polak y Ribiere utilizando como valores iniciales aquellos obtenidos en el ajuste previo.

El número máximo de iteraciones (**Max.Iter**) debe ser ingresado por el usuario (InfoStat tiene predefinido 1000) y en la salida se debe controlar que el algoritmo haya convergido naturalmente (es decir, que el número de iteraciones realizadas haya sido menor a Max.Iter). En el caso de no lograr convergencia (número de iteraciones igual a Max.Iter) se deberá re-ejecutar el procedimiento, preferentemente desde otro conjunto de valores iniciales y/o aumentando el valor de Max.Iter. Activando el campo **Resultados Iteraciones** se podrá visualizar los valores de la función objetivo en cada paso del proceso de optimización numérica seleccionado. Si el campo **Hoja Resultados** está activado InfoStat estimará el modelo que el usuario seleccionó en la solapa **Modelos** y presentará los resultados de esa estimación en la salida correspondiente. Si en el campo **Archivo ASCII** se ingresa un nombre, InfoStat guardará un archivo de texto con el nombre ingresado por el usuario.

Al estimar un modelo es posible guardar residuos (diferencia entre el valor observado y predicho por el modelo estimado), valores predichos por el modelo y los coeficientes de la función impulso respuesta en la tabla activa.

La función **impulso-respuesta** estima los coeficientes de la representación de Wold (Hamilton, 1994**)**. Se llama así a la representación gráfica de los $\psi_j's$ (coeficientes de la función impulso-respuesta) versus el índice *j*. Su interpretación es importante en estudios aplicados en Economía por cuanto describe los efectos de un shock instantáneo y único sobre la serie de tiempo en estudio.

La descomposición de Wold conceptualmente expresa que a toda serie de tiempo estacionaria se le asocia una representación infinita MA. El teorema de la descomposición de Wold (Hamilton, 1994) establece que si $E(y_t) = 0,\ y_t$ es un proceso de covarianza estacionario, entonces $y_t = \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j} + d_t$

donde $\{\Psi_j\}$ son constantes,

$\varepsilon_t = y_t - E(y_t \mid \Im_{t-1})$, i.e. $\varepsilon_t \sim MDS$ y tal que:

$$E\{\varepsilon_t \varepsilon_s\} = \begin{cases} \sigma^2 & s = t \\ 0 & s \neq t \end{cases},$$

$$E(\varepsilon_t y_{t-j}) = 0 \qquad \forall\ t, j > 0\text{, y}$$

$$E(\varepsilon_t d_s) = 0 \qquad \forall\ s \neq t.$$

Como herramientas para la validación del modelo identificado y estimado, el usuario puede realizar distintos tipos de operaciones sobre los residuos del modelo (**Validación sobre**

residuos). Si el modelo ajustado es bueno, los residuos debieran conformar una muestra de una distribución normal con media cero y la varianza constante y no poseer ningún tipo de correlación serial. Para analizar los residuos, InfoStat permite obtener los estadísticos MAD (mediana de los desvíos absolutos de los residuos respecto de su mediana), Rango (diferencia entre el valor mínimo y máximo de la serie de residuos), Kurtosis y Asimetría (debiera mostrar valores cercanos a tres para kurtosis y a cero para el coeficiente de asimetría), las funciones de autocorrelación y autocorrelación parcial de la serie de residuos (no debiera ajustar ningún modelo) y la prueba de normalidad propuesta por Jarque y Bera (Jarque y Bera, 1987) (valores *p* menores al nivel de significación sugieren el rechazo de la hipótesis nula de distribución normal). El estadístico de la prueba de Jarque y Bera es:

$$JB = \frac{T}{6}\left(A^2 + \frac{(K-3)^2}{4} \right) \overset{d}{\rightarrow} \chi^2_2$$

donde *T* representa el tamaño de la muestra, *A* es el coeficiente de asimetría definido como $A=\frac{1}{T}\frac{\sum e_i^3}{S^3}$, *K* es el coeficiente de kurtosis definido como $k=\frac{1}{T}\frac{\sum e_i^4}{S^4}$ y *S* es el desvío estándar.

Para el estudio de correlación serial de residuos, también se calcula el estadístico de Durbin-Watson, cuyo expresión es:

$$DW = \frac{\sum_{t=2}^{T}(\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T}\hat{e}_t^2}$$

donde $\hat{e}_t$ son los residuales del modelo ARMA(*p*,*q*) identificado y estimado. Puede probarse que 0<*DW*<4. Un valor de *DW* cercano a 0 indica autocorrelación positiva ($\rho > 0$). Un valor cercano a 4 indica una autocorrelación negativa ($\rho < 0$). Cuando *T* es grande, $DW \cong 2 - 2\hat{\rho}$, lo que sugiere que cuando $\hat{\rho}$ es próximo a cero (*i.e.* no autocorrelación) luego *DW* estará cercano a 2. La distribución del estadístico *DW* es no convencional y depende de las variables regresoras en el modelo ARMA(*p*,*q*). InfoStat deriva los valores *p* exactos para la prueba de la hipótesis de falta de correlación serial basada en el estadístico *DW*.

Para probar la hipótesis conjunta de que todos los coeficientes de autocorrelación son cero, InfoStat también obtiene los estadísticos de Box-Pierce y Ljung-Box (Pindyck y Rubinfeld, 1999). Estos estadísticos tienen la ventaja de no depender de las variables regresoras como sucede con el estadístico de Durbin-Watson, pudiendo mostrar valores de potencia sustancialmente mayores al estadístico *DW*. El estadístico de Box-Pierce (con distribución asintótica Chi cuadrado), es calculado como:

$$T\sum_{j=1}^{m}\hat{\rho}_j^2 \overset{d}{\rightarrow} \chi^2_{(m-p-q)}$$

donde *m* es el número de coeficientes involucrados en la prueba. Si el valor calculado del estadístico es mayor que el nivel crítico del 5%, podemos asegurar (con un nivel de significación de 0.05) que los coeficientes de autocorrelación verdaderos, ρ_1, ..., ρ_m no son todos iguales a cero. El estadístico de Ljung-Box (con distribución asintótica Chi cuadrado), es calculado como:

$$LB = T(T+2)\sum_{j=1}^{m}\frac{1}{T-j}\hat{\rho}_j^2 \overset{d}{\rightarrow} \chi^2_{(m-p-q)}$$

En la práctica, InfoStat toma $m = \min\left(\frac{T}{2}, 3\sqrt{T}\right)$. A medida que *T* se hace grande sucede que la distribución de estos estadísticos se aplana, $\chi^2_m \overset{d}{\rightarrow} U(0,\infty)$. Como solución a este problema, InfoStat también obtiene automáticamente la prueba LB estandarizada basada en el siguiente resultado:

$$\frac{LB-(m-p-q)}{\sqrt{2(m-p-q)}} \overset{d}{\rightarrow} N(0,1)$$

La subventana **Datos faltantes** puede ser utilizada para la predicción de valores faltantes o para detectar y medir efectos de la influencia de observaciones aberrantes. Si el campo **Predecir** no está activado, InfoStat calculará automáticamente los valores que encuentre en blanco en una celda de la serie, si está activado en cambio, el usuario puede ingresar la cantidad de casos que quiere predecir en el campo que se encuentra a la derecha de **Predecir**, al activarlo se hará visible una grilla donde se debe colocar los números de los casos a predecir. Ambas estrategias, predicción de valores en celdas vacías y celdas referenciadas por el usuario, pueden ser implementadas simultáneamente.

Si el campo **Hoja Resultados** está activado InfoStat estimará el modelo que el usuario seleccionó en la solapa **Modelos** y presentará los resultados de esa estimación en la salida correspondiente (ventana **Resultados**). La opción **Escribir grilla** cuando es activada permite agregar a la tabla activa una nueva columna, la cual contiene la serie original completada con los valores faltantes que se acaban de predecir. Opcionalmente, el usuario puede indicar el nombre de un archivo de texto en el campo **Archivo ASCII** de la subventana **Datos faltantes,** en el que se guardarán las predicciones de los datos faltantes.

InfoStat predice los valores faltantes a través del procedimiento basado en la esperanza de dicha observación condicionada a la información disponible. Este predictor lineal se construye postulando las restricciones propuestas por Alvarez *et al*. (1993). El procedimiento identifica los valores iniciales de las series hasta el primer dato faltante, luego predice todos los valores restantes (dato faltante más información disponible posterior al dato faltante). La predicción se realiza imponiendo restricciones de manera tal de minimizar el error de predicción de la información *a posteriori* del dato faltante (esta restricción implica que se predice con exactitud a los datos conocidos siguientes a los datos faltantes). Se demuestra que el predictor obtenido es el mejor predictor lineal insesgado en el sentido de ser aquel de menor error cuadrático medio de predicción si el modelo es conocido (Guerrero y Peña, 2000). No obstante es importante destacar que el modelo

verdadero rara vez es conocido. La valoración de estas predicciones cuando el modelo utilizado no corresponde a aquel asociado al generador de la serie, ha sido analizada por simulación (Smrekar, Robledo y Di Rienzo, 2001).

Los **Pronósticos** que se pueden obtener en esta subventana de InfoStat se basan en métodos de extrapolación e involucran la proyección de patrones o relaciones observadas en el pasado sobre el futuro. Los pronósticos son realizados a partir del modelo ARIMA estimado. El usuario debe elegir qué observaciones de la serie deberán ser usadas para emitir el pronóstico en los campos **Datos a usar**, **Primero** y **Ultimo**.

Si se desea pronosticar por ejemplo h=4 nuevos valores de la serie utilizando el modelo estimado a partir de las T observaciones, el usuario deberá ingresar el número 4 en el campo **Nro. pasos** y los números 1 y T en los campos **Primero** y **Ultimo**. Pronósticos a largo plazo (número de pasos alto) construidos a partir de modelos ARIMA son menos confiables que pronósticos a corto plazo.

La posibilidad de señalar los valores iniciales y finales de los datos utilizados para realizar el pronóstico permite realizar validaciones cruzadas de modelos ajustados en InfoStat. Así por ejemplo, si en el campo **Ultimo** se ingresa el valor T-h, InfoStat utilizará una serie de longitud T-h para estimar los parámetros del modelo y calculará automáticamente los errores cuadráticos de predicción (diferencia cuadrática entre el valor observado y el predicho a partir del modelo) para las últimas cuatro observaciones.

La reestimación del modelo a los fines predictivos se llama **Calibración**. En InfoStat, la calibración se puede realizar activando el campo **fija**, **recursiva** o **móvil**. En el campo **fija** los h valores pronosticados se construyen en base al modelo estimado con los datos indicados por el usuario en los campos **Primero** y **Ultimo**. En la calibración recursiva, el ancho de ventana se va actualizando (incrementando) cada vez que se pronostica un nuevo valor. La actualización se realiza agregando a la serie de datos a usar para la estimación un nuevo dato disponible. En la calibración móvil, la ventana se mantiene de ancho fijo, de forma tal que se usa para pronosticar la observación T+h los datos desde h-1 hasta T+h-1.

Para todas las predicciones realizadas, InfoStat permite obtener las bandas de predicción de coeficiente (1-α)100%, ingresado por el usuario en el campo **Bandas al:**. La opción **Escribir grilla** cuando es activada dentro del submenú **Pronósticos** permite agregar a la tabla activa una nueva columna la cual contiene la serie original completada con los valores que se acaban de pronosticar. Opcionalmente, el usuario puede indicar el nombre de un archivo de texto en el campo **Archivo ASCII** de la subventana **Pronósticos** en el que se guardarán los pronósticos y los errores de predicción en el caso de que éstos puedan ser calculados por InfoStat. Si el campo **Hoja Resultados** está activado, InfoStat realizará los pronósticos que se solicitaron y presentará los resultados de esas proyecciones en la ventana **Resultados**. Se entiende que el objetivo del usuario al ingresar a esta subventana es el de generar pronósticos o proyecciones que en algún sentido sean óptimos.

En caso que los pronósticos sean grabados en la tabla activa, estos podrán ser graficados desde el panel **Gráficos**.

Una función de pérdida es una función que permite describir cuán costoso resultará si el pronóstico se aleja una cierta magnitud o distancia del verdadero valor. Como función de pérdida, InfoStat reporta automáticamente el **Error cuadrático medio de predicción**

(ECMP), el cual representa la función objetivo que debe ser minimizada en el proceso de calibración. Siguiendo a Hamilton (1994), si denotamos por $Y^*_{t+k|t}$ el pronóstico de Y_{t+k} en base a un conjunto de variables Y_t, observadas en la fecha o momento *t*, luego el ECMP es:

ECMP = $E(Y_{t+1} - Y^*_{t+1|t})$. Así, es posible demostrar (ver Hamilton, 1994) que el pronóstico de un paso adelante que minimiza el ECMP es tal que $\hat{y}_{t+1} = E(y_{t+1} \mid y_t \subset \Im_t)$ y el pronóstico *k* pasos adelante es tal que $\hat{y}_{t+k} = E(y_{t+k} \mid y_t \subset \Im_t)$.

Para la construcción de intervalos de confianza, InfoStat calcula el error de pronóstico de *k* pasos adelante como**:**

$$\begin{aligned} e_{T+k} &= y_{T+k} - \hat{y}_{T+k} \\ &= \psi_0 \varepsilon_{T+k} + \psi_1 \varepsilon_{T+k-1} + \cdots + \psi_{k-1} \varepsilon_{T+1} \end{aligned}$$

con varianza del error de pronóstico expresada como:

$$\begin{aligned} V(e_{T+k}) &= E(y_{T+k} - \hat{y}_{T+k})^2 \\ &= (\psi_0^2 + \psi_1^2 + \cdots + \psi_{k-1}^2)\sigma_\varepsilon^2 \\ &= (1 + \sum_{j=1}^{k-1} \psi_j)\sigma_\varepsilon^2. \end{aligned}$$

La estimación se basa en la suma de residuales al cuadrado obtenida después de que se han conseguido las estimaciones de los parámetros del modelo ARMA especificado, esto es:

$\hat{\sigma}^2_\varepsilon = \dfrac{\sum_{t=1}^{T} \hat{\varepsilon}_t}{T - p - q}$ en caso que los parámetros se hayan estimado en base a la función de verosimilitud exacta o $\hat{\sigma}^2_\varepsilon = \dfrac{\sum_{t=p+1}^{T} \hat{\varepsilon}_t}{T - 2p - q}$ si se usó la función de verosimilitud condicional (implementado en InfoStat).

En base a esto, conociendo la forma de la varianza del error de predicción y asumiendo normalidad de los términos de error, sigue que el intervalo de confianza $(1-\alpha)\times 100\%$, se puede aproximar por medio de:

$$\hat{y}_{t+k} \pm z_{1-\frac{\alpha}{2}} \left(1 + \sum_{j=1}^{k-1} \hat{\psi}_j^2 \right)^{1/2} \hat{\sigma}_\varepsilon.$$

Ejemplo 46: Sobre una serie de 100 observaciones simulada en InfoStat especificando un modelo AR(2), se realizó el proceso de estimación de un modelo AR(2) utilizando el

algoritmo de Nelder y Mead, indicando los valores 1 y T-h=96 en los campos Primero y Ultimo de la subventana pronóstico (calibración fija), y Nro. Pasos h= 4. Los resultados se presentan en la Tabla 67.

Tabla 67: Resultados del proceso de estimación, validación y pronóstico de un modelo AR(2) sobre una serie AR(2).

Estimación maximoverosímil ARIMA
Algoritmo numérico de optimización: Nelder&Mead
Información general

Serie	Nro.Obs.	Media	Varianza muestral	Desvío Estándar
Serie3(1)	96	-0.11	2.97	1.72

Información sobre pronósticos realizados

Obs	Pronóstico	LIP(95)	LSP(95)
97.00	-0.01	-1.96	1.94
98.00	-0.47	-2.49	1.55
99.00	-0.12	-2.52	2.28
100.00	-0.29	-2.79	2.20
ECMP:	0.86	-------	

Resultados de la Estimación

Parámetro	Estimación	Error Std	t-val	Valor p
Cte	0.02	0.10	0.16	0.8738
=> Mu_Y	0.12	0.77	0.16	0.8747
AR(1)	0.28	0.08	3.34	0.0012
AR(2)	0.59	0.08	7.08	<0.0001

Medidas Resumen y Validación

Estadístico	Valor Observado	Valor p
Verosimilitud	-132.35	
CMResidual	0.99	
R^2	0.67	
R^2 Corregido	0.67	
Akaike IC:	0.05	
Schwarz IC:	0.13	
Hannan-Quinn IC:	0.08	

Iter.para Converger:	139	
MAD	0.68	
Rango Residuos	4.81	
Asimetría Residuos	-0.11	
Kurtosis Residuos	2.68	
Normal.(Jarque-Bera)	0.62	0.7347

Función autocorrelación: r(k) (se muestran sólo los 5 primeros lags)

Lag	Coef	se r(k)	t-val	Valor p	Signif
1	-0.01	0.10	-0.11	0.9110	
2	0.01	0.10	0.12	0.9076	
3	0.09	0.10	0.85	0.3964	
4	0.13	0.10	1.29	0.2010	
5	0.16	0.11	1.52	0.1327	

Función autocorrelación parcial: phi(k)
(se muestran sólo los 5 primeros lags)

Lag	Coef	se r(k)	t-val	Valor p	Signif
1	-0.01	0.10	-0.11	0.9110	
2	0.01	0.10	0.12	0.9086	
3	0.09	0.10	0.86	0.3947	
4	0.14	0.10	1.33	0.1875	
5	0.17	0.10	1.64	0.1073	

En la salida en **Información sobre pronósticos realizados**, se puede visualizar el bloque de información general sobre la serie y sobre los pronósticos realizados para las observaciones 97 a 100, con sus respectivos límites para el intervalo de predicción (LIP y LSP). El error cuadrático medio de la predicción es 0.86.

En la tabla **Resultados de la Estimación**, se muestran las estimaciones de los parámetros con sus respectivos errores estándares, estadístico *T* y valores *p* calculados bajo el supuesto de normalidad. El término Cte=0.02 hace referencia a la constante del polinomio AR, el que se relaciona con la esperanza de los datos de la forma $Cte/(1+\sum_{j=1}^{p} AR(j))$ y que en la tabla se indica con ==> Mu_Y y para el ejemplo su valor es 0.12. Este parámetro no es estadísticamente significativo ya que la serie fue originada a partir de un proceso con esperanza cero.

Los coeficientes AR son estadísticamente significativos con valores estimados de 0.28 y 0.59 (la serie bajo análisis fue generada por un proceso AR(2) con coeficientes 0.3 y 0.5).

La tabla **Medidas Resumen** y **Validación** muestra valores de estadísticos que permiten suponer un buen ajuste. La prueba de normalidad indica que no existen evidencias para rechazar el supuesto de normalidad de la serie de los residuos. Por último se presentan las estimaciones de la función de autocorrelación y la función de autocorrelación parcial de los residuos las que sugieren que no existe correlación entre los mismos (recordar que fueron generados de forma independiente).

Suavizados y ajustes

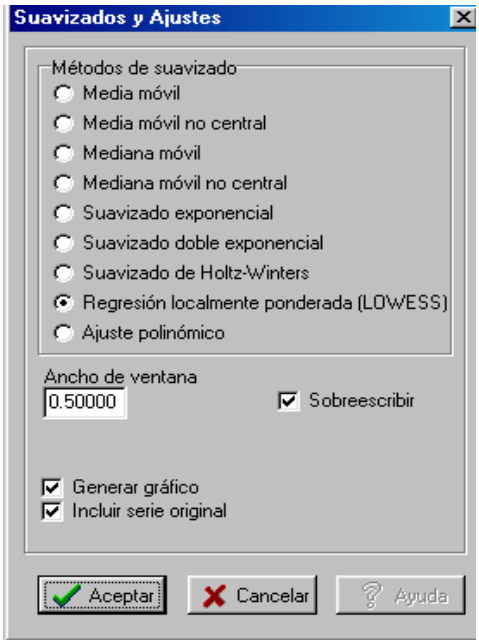
MENÚ⇒ ESTADÍSTICAS ⇒ SUAVIZADOS Y AJUSTES, permite utilizar herramientas para describir tendencias en una variable dependiente (Y) como función de una o más variables regresoras (X). Las técnicas de suavizado implementadas no necesitan de la especificación de un modelo; son útiles para filtrar variación en el diagrama de dispersión de una variable dependiente que entorpece la visualización de tendencias de dicha variable respecto a X.

InfoStat provee suavizados del tipo lineal, es decir, en la serie suavizada cada elemento es una combinación lineal del elemento correspondiente en la serie original. Si bien la serie suavizada provee un estimador sesgado de la tendencia, puede representar una sustancial ganancia para la interpretación debido a la menor varianza de la serie suavizada. Para elegir entre varios suavizados lineales usualmente se comparan los errores cuadráticos medios (sesgo al cuadrado más varianza). Mientras menor sea dicho valor, mejor será la estrategia de suavizado.

Las distintas técnicas de suavizados se basan en la elección de una función para ser aplicada en el vecindario de cada observación. El vecindario se define como el conjunto de observaciones (X_i,Y_i) que están por arriba y por debajo (o antes y después) de cada valor de la variable regresora (excepto X_{max} y X_{min}).

Cuando se utilizan suavizados se recomienda graficar simultáneamente las series de observaciones suavizadas y originales en función de la variable regresora o una secuencia

ordenada para corroborar que la técnica de suavizado no esté proveyendo una señal falsa. InfoStat permite obtener automáticamente estos gráficos desde el menú SUAVIZADOS Y AJUSTES.



Algunas técnicas de ajuste, como son el *ajuste polinómico* y el *ajuste estacional*, pueden ser interpretadas como una forma especial de suavizado. InfoStat permite ajustar polinomios de alto orden para eliminar fluctuaciones irregulares de alta frecuencia.

En la ventana **Suavizados** se elige la/s **Variables dependientes** (variables cuya tendencia desea maximizar por suavizado) y en caso de existir la variable a considerar en el eje X se debe elegir **Ordenamiento/Regresoras (optativa)**. Si esta variable no se especifica, la secuencia será ordenada tal como han sido ingresados los datos de Y en la tabla (desde el caso 1 al caso *T*, siendo *T* el número de casos). Al **Aceptar** se habilita una segunda ventana que permite seleccionar la técnica de suavizado y si esta lo requiere se deberá tomar una decisión sobre el ancho de ventana, valor éste que determina el número de observaciones que serán consideradas como pertenecientes al vecindario de cada observación. Para ello el usuario debe remitirse al campo **Ancho de ventana**. Si se utiliza reiteradamente suavizados para una misma variable, se generará por cada suavizado una nueva columna en la tabla de datos. Para evitar la proliferación de columnas se deberá activar el campo **Sobreescribir**. También se puede solicitar **Generar gráficos** para graficar la serie suavizada, **Incluir serie original**, para suponer en el mismo gráfico los datos originales y suavizados y **Particiones en el mismo gráfico** para representar dos o más series suavizadas.

Técnicas de suavizado

Media móvil y Media móvil no centrada: los suavizados del tipo **media móvil** se basan en el uso de valores Y, correspondientes a una ventana móvil de puntos en X, para estimar tendencias en Y. Para una media móvil de ancho de ventana *n*, el valor t-ésimo de la serie suavizada es la media aritmética de las observaciones Y_t, Y_{t-1}, Y_{t-2}, ..., Y_{t-n+1} de la serie original. Es decir, en el suavizado no centrado sólo se usan como vecinos valores anteriores al valor que está siendo transformado. Si la serie es una serie de tiempo (secuencia ordenada en el tiempo) se podría decir que sólo se usan valores del pasado. InfoStat también permite obtener medias móviles (**Media móvil**) donde el vecindario de cada dato está conformado por los vecinos más cercanos y no necesariamente aquellos anteriores. En un promedio móvil de ancho de ventana 5, se reemplaza cada valor Y_t por $\overline{Y}_t$, siendo $\overline{Y}_t$ la media de las 5 observaciones más cercanas a Y_t. Si la variable X asume valores equidistantes, la media

móvil de ventana 5 para la observación Y_t (excepto para Y_t extremos de la serie) se calculará a partir de las observaciones $Y_{t-2}, Y_{t-1}, Y_t, Y_{t+1}, Y_{t+2}$

Mediana móvil y Mediana móvil no centrada: los suavizados del tipo mediana móvil se obtienen de forma análoga a los del tipo media móvil (ver Media móvil), pero la función aplicada para obtener el valor suavizado, Y_t, es la mediana de los valores del vecindario en lugar de la media.

Suavizado exponencial: corresponde al uso del modelo de medias móviles (ver Media móvil no centrada) ponderado exponencialmente de manera tal que se asignan pesos mayores a los valores de Y más cercanos a Y_t. El elemento t-ésimo valor de la serie suavizada exponencialmente es:

$$\overline{Y}_t = aY_t + a(1-a)Y_{t-1} + a(1-a)^2 Y_{t-2} + \ldots$$

donde la sumatoria se extiende retrospectivamente en toda la serie. La forma recursiva para obtener $\overline{Y}_t$, usada por InfoStat es:

$$\overline{Y}_t = aY_t + (1-a)\overline{Y}_{t-1}$$

Valores pequeños de *a* determinan un mayor suavizado.

Suavizado doble exponencial: datos suavizados a partir del suavizado exponencial (ver Suavizado exponencial), son suavizados nuevamente con un filtro exponencial. Esta acción permite lograr un mayor suavizado sin dar tanto peso a observaciones individuales anteriores (del pasado). El elemento *t*-ésimo de la serie suavizada doble exponencialmente es:

$$\overline{\overline{Y}}_t = a\overline{Y}_t + (1-a)\overline{\overline{Y}}_{t-1}$$

Suavizado de Holtz-Winters: se obtiene aplicando la función de suavizado exponencial simple con la adición de cambios medios en tendencia (incremento o disminución). El elemento t-ésimo de la serie suavizada se obtiene a partir del siguiente sistema de ecuaciones:

$$\overline{Y}_{tt} = aY_t + (1-a)(\overline{Y}_{t-1} + r_{t-1})$$
$$r_t = b(\overline{Y}_{tt} - \overline{Y}_{t-1}) + (1-b)r_{t-1}$$

donde *a* y *b* son pesos (parámetros del suavizado) que asumen valores en el intervalo [0,1] y r_t una serie suavizada representando la tasa promedio de cambio en la serie $\overline{Y}_t$.

Nota: Para aplicaciones específicas de estos suavizados para pronósticos en series de tiempo (ver Series de tiempo).

Gráficos

El componente gráfico de InfoStat consta de dos ventanas de interfase con el usuario. Una ventana de **Herramientas Gráficas** y otra llamada **Gráficos** que muestra los gráficos propiamente dichos. Estas ventanas se actualizan para reflejar las características del gráfico que se está visualizando.

Un gráfico de InfoStat es una colección de series gráficas, un sistema de coordenadas, un título, una colección de elementos de texto y una leyenda. Un gráfico tiene un tamaño por defecto que el usuario puede modificar cambiando el tamaño de la ventana gráfica. Sobre un gráfico activo pueden seleccionarse opciones asociadas al botón derecho del ratón.

Cada serie gráfica es un conjunto de elementos gráficos de un único tipo (puntos, sectores de una torta, rectángulos, líneas, etc.). Las series gráficas tienen atributos que el usuario puede modificar como: nombre, color, forma de sus elementos y orden en que se grafican las series (si el gráfico tiene más de una), etc. Existe un menú de opciones para una serie gráfica que está asociado al botón derecho del ratón y es aplicable a las series seleccionadas desde la solapa **Series** de la ventana de **Herramientas gráficas**.

Los elementos gráficos de una serie tienen un cuerpo (que es el que se colorea y cambia de forma), un borde que delimita al cuerpo, brazos (superior e inferior) y terminaciones de los brazos.

El sistema de coordenadas tiene un eje X y uno o varios ejes Y (eventualmente uno para cada una de las series), cada eje tiene su leyenda, su escala y tipo, que puede ser numérico o categórico. Las propiedades más importantes, de un eje que representa escalas numéricas que el usuario puede modificar son: el mínimo y máximo, el número de decimales, el número de divisiones de la escala numérica, el espesor y el color. Para ejes X categóricos, el usuario puede además cambiar la secuencia de las categorías y sus nombres. En el caso de tener un número grande de categorías, es posible presentarlas en una secuencia en la que se alterna su posición respecto del eje facilitando su visualización o mostrar sólo un subconjunto de ellas. Los ejes pueden incluir líneas de corte cuyo espesor y color se pueden modificar.

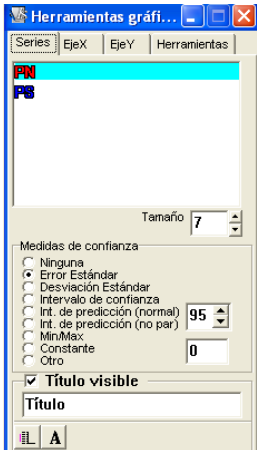
El título, las leyendas de los ejes y los textos que aparecen en el gráfico, ya sea porque el usuario los agregó o porque son mostrados automáticamente por InfoStat, son objetos que pueden moverse, editarse y modificarse estéticamente. En la ventana **Herramientas gráficas** y en las correspondientes solapas (**Series**, **EjeX**, **EjeY**, **Herramientas**), se encontrarán las opciones para ello.

La leyenda es un objeto ligado a las series y el número de entradas en una leyenda se corresponde con el número de series cuyo atributo de leyenda esté visible. El nombre, color y forma de elemento gráfico asociado y el orden en que se presentan los ítems de la leyenda se corresponde con el nombre y el orden de las series. La posición de la serie puede modificarse en cuyo caso se modificará la posición de la leyenda. Para acceder a las opciones de la leyenda seleccionar la misma con el ratón y activar el botón derecho.

A continuación se presentan las utilidades disponibles para modificar y ajustar un gráfico. La presentación comienza con la descripción de ventana de **Herramientas gráficas** y luego con la ventana de **Gráficos**.

Herramientas Gráficas

Esta ventana aparece siempre junto a la ventana **Gráficos**. Si se ha cerrado, se activa nuevamente accionando el botón izquierdo del ratón cuando el cursor está posicionado sobre un gráfico. Tiene cuatro solapas: **Series**, **Eje X**, **Eje Y** y **Herramientas.**



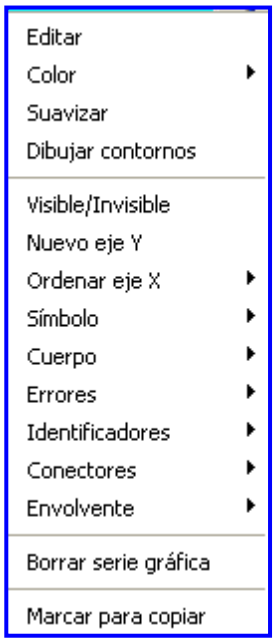
Sección Lista de series

Sección Opciones

Sección Título

Solapa Series

Esta solapa activa un panel organizado en tres partes. La sección superior contiene una lista de las series incluidas en el gráfico. La sección media presenta opciones aplicables a las series seleccionadas de la lista anterior. Su contenido depende del tipo gráfico de las series seleccionadas. La sección inferior contiene un campo de edición del título del gráfico y algunos botones que activan opciones sobre el mismo (cambiar la tipografía, reponerlo en su posición por defecto y hacerlo visible o invisible).



Para modificar las propiedades de una serie gráfica, como su color, la serie debe estar seleccionada. Una serie se selecciona accionando el botón izquierdo del ratón con el cursor apuntando al nombre de la serie que aparece en la lista de series del gráfico. Varias series pueden seleccionarse simultáneamente arrastrando el ratón sobre ellas con el botón izquierdo apretado. Si una o varias series están seleccionadas aparecen resaltadas con color un poco más oscuro de fondo. Una serie seleccionada puede moverse para alterar el orden de graficación manteniendo apretada la tecla <Ctrl> y accionando las teclas de movimiento (flechas) hacia arriba o abajo. El cambio de posición de la serie se refleja en el orden de las entradas de la leyenda.

Una o más series seleccionadas y el cursor sobre alguna de ellas habilita, presionando el botón derecho del ratón, un menú de propiedades de la serie que se pueden modificar. Los ítems presentados en este menú dependerán del tipo de serie gráfica seleccionada. A su vez, aparecerán activos o no, dependiendo de las características particulares de la serie.

Editar abre una ventana de diálogo para cambiar el nombre de la serie. Esta ventana también se activa haciendo doble *click* sobre el nombre de la serie o apretando la tecla

<Enter> cuando la serie está seleccionada. Cuando el nombre de una serie se modifica, también se modifica su entrada en la leyenda de la figura. Esta es la forma de editar el contenido de una leyenda.

El ítem **Color** despliega un submenú con una lista de los colores disponibles. Al seleccionar uno de ellos, el cuerpo de los elementos gráficos de la serie cambian de color y hace lo propio el nombre con el que aparece la serie en la lista.

El ítem **Suavizar** se habilitará por ejemplo para diagramas de dispersión o diagramas de puntos. Al activarlo, InfoStat genera una *serie suavizada* que se agrega al gráfico realizado y que aparece también en la ventana de las **Series**. Esta serie tiene opciones propias que permiten modificarla (ver Suavizados y ajustes).

El ítem **Dibujar contornos** se habilitará cuando las series gráficas activadas sean diagramas de dispersión o diagramas de puntos entre otras. Al activarlo, InfoStat genera una *serie contorno* que se agrega al gráfico y que al igual que la serie suavizada también se agrega en la ventana **Series**. Esta serie tiene opciones propias que permiten modificarla y que aparecerán en forma automática al pintar la serie suavizada.

El ítem **Visible/Invisible**, hace visible o invisible las series seleccionadas. Si una serie está invisible, su nombre aparecerá, en la ventana de las **Series**, con un sombreado diferente, más claro.

Nuevo eje Y, agrega un nuevo eje Y para cada una de las series seleccionadas. Los nuevos ejes aparecerán alternadamente a la derecha y a la izquierda del gráfico y recibirán un número correlativo comenzando en 1 (uno). El eje original del gráfico es el eje 0 (cero). La posición de un eje Y puede modificarse (como se indicará más adelante). Los ejes Y "adicionales" pueden borrarse. En este último caso el eje de referencia para la serie cuya escala se reflejaba en el eje borrado pasa a ser nuevamente el eje 0 (cero).

El ítem **Ordenar eje X** permite ordenar los valores de la variable en X en forma **Alfabética**, **Según valores crecientes de Y** y **Según valores decrecientes de Y**, si su naturaleza es categórica.

El ítem **Símbolo** permite cambiar la forma del *cuerpo* de los elementos gráficos que conforman la serie. Las formas disponibles aparecen en una lista como un submenú de este ítem. La forma de los elementos gráficos no puede modificarse en todos los tipos de series gráficas, en cuyo caso el ítem aparece deshabilitado.

El ítem **Cuerpo**, permite optar por las siguientes opciones: **Relleno/Vacío**, **Con/Sin borde**. Por ejemplo si el cuerpo de los puntos es sólido al activar la primera opción se verán sólo los contornos del mismo. Si se vuelve a activar esta opción se rellenará nuevamente. Lo mismo sucede con el borde del símbolo si se invoca la opción **Con/Sin borde**.

El ítem **Errores**, contiene opciones para hacer visibles o invisibles los errores asociados a la medida representada en un elemento gráfico. Los errores, ya sean errores estándares, desvíos estándares, intervalos de confianza o intervalos de predicción, se representan en forma simbólica con un segmento cuya longitud se relaciona con la magnitud del mismo. Este ítem no está habilitado para todos los tipos de series gráficas.

El ítem **Identificadores** activa y desactiva la visualización de rótulos que identifican a los elementos de las series gráficas. Estos rótulos pueden contener la coordenada del elemento gráfico en el eje Y, su coordenada en el eje X, las coordenadas (X,Y), el número de caso que el ítem representa (si es pertinente) o contener un rótulo arbitrario. Algunas técnicas gráficas como los diagramas de dispersión, permiten asociar una columna de la tabla de datos con los rótulos que se quieren mostrar.

Bajo el ítem **Identificadores**, también se incluyen opciones para modificar el aspecto de los rótulos. Cualquier rotulo visible en un gráfico puede ser movido o editado dentro del gráfico. Cuando alguna de estas dos cosas ocurre, el rotulo queda "tocado". Bajo esta condición el rótulo queda anclado al lugar al que se lo movió o permanece anclado en el lugar donde fue editado. Luego, si el gráfico es redimensionado, o se cambian las escalas de los ejes, estos "rótulos modificados" se "mueven" de manera ligeramente diferente del resto.

Para volverlos a su posición y contenido por defecto, se debe seleccionar el rótulo dentro del gráfico apretando el botón izquierdo del ratón con el cursor apuntándolo y luego accionar el ítem *Reset* que aparece al final de un menú desplegable que se activa con el botón derecho del ratón.

El ítem **Conectores** se utiliza para hacer visible o invisibles líneas que unen los cuerpos de los elementos gráficos de una serie (conectores). Bajo este menú también hay opciones para cambiar el color, espesor, relleno (trama) y plano de visualización de los conectores.

El ítem **Envolventes** hace visible o invisible una clase especial de conectores que une los extremos de la barras de error y permiten crear bandas de confianza o predicción. Al igual que el ítem **Conectores**, este menú tiene opciones para mejorar el aspecto de los conectores como el color y el espesor.

El ítem **Borrar serie gráfica** permite eliminar una o más series seleccionadas previamente.

El ítem **Marcar para copiar** se utiliza para copiar series de un gráfico y superponerlas en otro a los fines de observar el comportamiento de dos variables. Se dará un ejemplo a continuación.

Ejemplo 47: Los datos del archivo Base1.idb corresponden a la Encuesta Permanente de Hogares realizada por el INDEC en el primer trimestre del año 2006. Utilizando el archivo Base1.idb se obtuvo el siguiente gráfico:

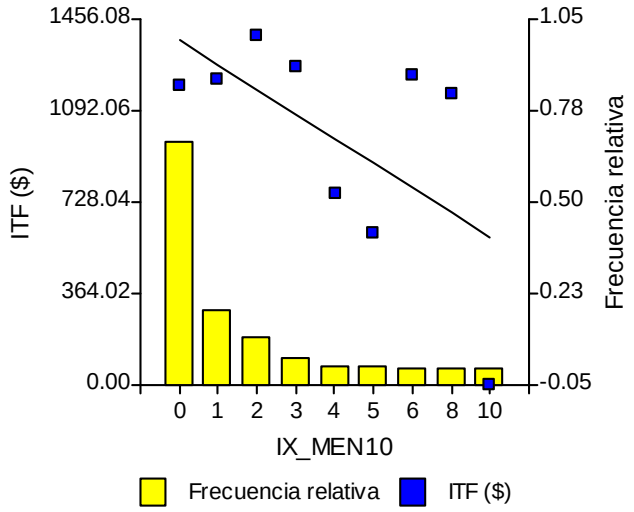


Figura 42: Gráfico de Barras para cantidad de miembros del hogar menores de 10 años (IX_MEN 10) y Gráfico de Puntos para la variable ingreso total familiar en pesos (ITF) versus cantidad de miembros del hogar menores de 10 años (IX_MEN 10). Archivo Base1.idb.

Esta figura se obtuvo haciendo un gráfico de barras con la frecuencia relativa para "IX_MEN10" como criterio de clasificación y poniendo "Caso" como variable a graficar. Luego de aceptar, en el menú desplegable que tiene por defecto "Media" seleccionar Frecuencia Relativa. Luego se realizó un gráfico de puntos para "ITF" versus "IX_MEN10". Sobre este último gráfico seleccionar la **Serie** (solapa Series) y con el botón derecho seleccionar **Marcar para copiar,** finalmente sobre el gráfico de barras accionar el botón derecho de ratón y elegir **Agregar series marcadas para mover**. También se agregó un eje adicional para la serie que faltaba (ITF) seleccionando la **Serie** y con el botón derecho del ratón activando **Nuevo eje y**. En el gráfico se puede adicionar una leyenda (sobre el gráfico apretar botón derecho y seleccionar **Mostrar leyenda**) para asociar los símbolos y colores con las variables analizadas, además, sobre la leyenda presionando el botón derecho del ratón en el menú que aparece se eligió **Borde**, **Invisible**. Por último sobre la serie ITF, activar la opción **Suavizado** y luego sobre esta nueva serie generada y en la misma ventana de la solapa **Series**, activar **Poli** y en la casilla adyacente **1**, actualizando posteriormente (botón **Actualizar**). Esta serie suavizada fue borrada de la leyenda (marcar la serie suavizada y activar **Leyenda visible/invisible**).

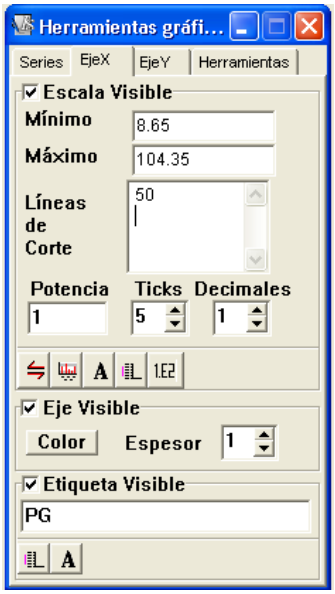
Solapa Eje X

Esta solapa presenta un panel que consta de tres secciones. La primera controla atributos generales de la escala. La segunda controla el espesor y el color del eje. La tercera exhibe la leyenda y facilita su edición, la modificación de sus atributos tipográficos y eventualmente la reposición de su ubicación por defecto.

La sección de atributos de escala contiene campos de edición para los valores mínimo, máximo del eje y una lista donde, opcionalmente, se pueden especificar puntos del eje X para trazar líneas de corte. InfoStat trazará líneas perpendiculares al eje X en los puntos

indicados anteriormente. El número y posición de estas líneas, así como su color y espesor, son controlados por el usuario.

El número de divisiones de la escala (*ticks*) y el número de decimales que muestra la escala pueden modificarse. Debe notarse que, cuando se incluyen identificadores que contienen la coordenada del eje X, el número de decimales con que aparece este valor se corresponde con el número de decimales especificados para la escala del eje.



Si la escala del eje X es numérica y el mínimo es mayor que 0, aparece un campo de edición adicional rotulado como **potencia.** Por defecto este campo de edición contiene un 1. Si se introduce en este campo un valor (α), InfoStat graficará en escala X^{α} . Cuando α=0, InfoStat utiliza escala logarítmica para los valores del eje.

La barra de botones de comando, que está en la base de la sección de atributos de la escala, permite de izquierda a derecha lo siguiente: invertir la escala (de menor a mayor o viceversa), especificar si los rótulos asociados a las divisiones de la escala se exhiben en una o dos líneas alternantes para facilitar su lectura, cambiar la tipografía de los mismos, volver a la posición original todos los rótulos de la escala y visualizar la escala en formato decimal o exponencial.

Cuando en el eje X se representa una variable categórica los dispositivos que permiten controlar el número de *ticks* y el número de decimales desaparecen. Asimismo los campos de edición del mínimo y máximo son sustituidos por una lista de las categorías representadas en el eje. En esta lista (que aparece en la subventana de la solapa **EjeX**), el nombre de las categorías puede ser editado (doble *click* o <Enter>) y su posición alterada (tecla <Ctrl>+flecha de dirección hacia arriba o abajo). Si la posición de una categoría se modifica, esto se reflejará en el gráfico.

Si el número de categorías es grande, como puede ocurrir cuando se grafican series temporales, puede ser imposible leer todos los rótulos correspondientes a las divisiones de la escala debido a su superposición. Para evitar este inconveniente se pueden disponer los rótulos en doble fila. Si esta opción no es suficiente para resolver las superposiciones, existe un dispositivo en la barra de botones de comando (aparece sólo cuando la variable en X es categórica) que permite especificar el número de categorías que deben contarse, a partir de la primer categoría visible, para encontrar la próxima visible (por defecto su valor es uno y en este caso todas las categorías son visibles).

Solapa Eje Y

Esta solapa es similar a la descripta para el eje X. Sin embargo ésta se actualiza con el eje Y seleccionado ya que puede haber tantos ejes Y como series en el gráfico. Un eje se selecciona presionando el botón izquierdo del ratón cuando el cursor apunta al eje deseado.

El número del “eje activo” se muestra en el borde superior derecho del marco que delimita la sección de atributos de la escala.

A los campos de edición del mínimo y el máximo, se agrega un campo llamado **Base** en el que se puede escribir un valor de referencia para lograr un gráfico con barras que se proyectan por encima de este valor si la diferencia entre la media y él es positiva y por debajo si es negativa (ver ejemplo en la sección Gráfico de barras). El campo de edición de la base está protegido y cualquier modificación no tendrá efecto mientras la protección esté activada. Para anular la protección, se debe desactivar la casilla que se encuentra a la izquierda del campo, haciendo *click* con el cursor posicionado sobre ésta, luego escribir el valor para la base.

Las líneas de corte en el eje Y son equivalentes a las del eje X, pero cada eje Y puede tener sus propias líneas de corte. La barra de botones en la base de la sección de atributos de la escala es similar a la descripta para la solapa del eje X pero incluye: un botón de comando con el ícono de un recipiente de basura que al activarlo elimina el eje seleccionado. Esta acción es aplicable cuando se tienen por lo menos dos ejes ya que no se puede eliminar el eje 0 (siempre debe quedar un eje de referencia, aunque puede hacerse invisible). Botones adicionales permiten reposicionar el eje seleccionado a izquierda o derecha del gráfico (botón **Invertir la posición del eje**) y moverlo punto a punto para separarlo de los elementos gráficos (botones con flechas izquierda y derecha) cuyas coordenadas en el eje X hacen que, eventualmente, el cuerpo de los elementos gráficos se superponga con el eje Y.

Al igual que el eje X, los ejes Y tienen un campo de edición **Potencia** que se aplica de manera idéntica a la descripta para el eje X.

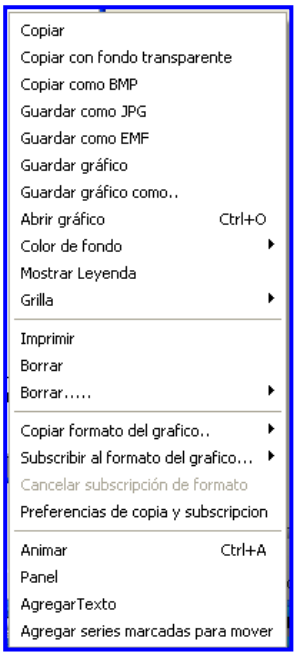
Solapa Herramientas

Botón Texto

En esta solapa hay un botón con el rótulo **Texto** que si se acciona permite insertar líneas de textos en lugares arbitrarios del gráfico. Cuando esta herramienta está activada, el cursor cambia de forma cuando se desliza sobre grafico indicando que está disponible para iniciar un “punto de inserción” de texto. Al hacer *click* sobre el gráfico se crea un campo de edición que permite insertar el texto deseado. La tipografía que por defecto utiliza esta herramienta es una tipografía global para todas las inserciones de texto de la ventana gráfica. Esta característica global se controla, mientras la ventana activa es gráfica, accionando el botón identificado con la letra **A** de la barra de botones de comando de la ventana principal de InfoStat. Si se cambia la tipografía global todos los textos insertados del gráfico activo, se modificarán de acuerdo a la elección tipográfica. No obstante, los atributos tipográficos de cada texto pueden modificarse individualmente, seleccionando el texto y activando el ítem de **tipografía** del menú que se despliega apretando el botón derecho del ratón.

Una forma rápida de agrandar o disminuir la tipografía de un gráfico es utilizando Ctrl. ↑ y Ctrl. ↓ sobre el gráfico.

Ventana Gráficos



En esta ventana, InfoStat almacena todos los gráficos que se realicen en una sesión de trabajo. Los gráficos son numerados consecutivamente a partir de cero. Haciendo *click* sobre las solapas numeradas que aparecen en la base de la ventana gráfica se accede al gráfico requerido. También se pueden escoger lo gráficos creados en esta ventana con el comando rápido CTRL+Flecha hacia la izquierda o CTRL+Flecha hacia la derecha. Asimismo, CTRL+Fin y CTRL+Inicio con las flechas izquierda y derecha permiten ir al último y primer grafico creados respectivamente. Todas las acciones que se pueden realizar sobre esta ventana están ligadas a un menú que se despliega con el botón derecho del ratón cuando el cursor está posicionado sobre la ventana de gráficos. Las acciones son: **Copiar**, para copiar el gráfico activo al portapapeles (*clipboard*) para ser transportado a un procesador de textos; **Copiar como BMP**, copia el gráfico con formato Bitmap (este formato es reconocido universalmente por cualquier procesador de texto y está destinado a evitar problemas de incompatibilidad con el formato Windows Meta Archivo Mejorado que es el que InfoStat usa por defecto cuando copia un gráfico en el portapapeles); **Guardar como JPG** y **Guardar como EMF** abre una ventana de diálogo que permiten guardar el gráfico activo en estos formatos. Por defecto InfoStat guardará el gráfico con el nombre del archivo a partir del cual se generó más el número correspondiente en la solapa de gráficos precedido por subguión. La opción **Guardar gráfico** abre una ventana de diálogo que permiten guardar todos los gráficos construidos en el formato de InfoStat (genera un archivo con extensión .IGB, que son reconocidos por InfoStat); **Abrir gráfico** permite abrir los gráficos guardados como IGB en la ventana **Gráficos** de InfoStat**; Color de fondo**, permite cambiar el color del fondo que por defecto es blanco; **Mostrar leyenda**, muestra u oculta las leyendas; **Grilla**, activa y desactiva una grilla que se corresponde con las subdivisiones de la escala de los ejes X e Y (un submenú facilita el retoque del aspecto); **Imprimir**, abre una ventana de diálogo para imprimir el gráfico activo; **Borrar**, borra el gráfico activo y **Borrar...**, da opciones para borrar hacia delante o atrás o borrar todos los gráficos creados. Además hay varios ítems de menú relacionados con la copia o suscripción de formatos que se detallan mas adelante. La opción **Animar** permite mostrar en forma alternada todos los gráficos creados hasta el momento y se puede activar y desactivar con el comando rápido CTRL+A. La opción **Panel** genera una copia del gráfico activo que permite mostrar los gráficos construidos hasta el momento en forma simultánea. Estos gráficos dispuestos en panel siguen ligados a la ventana grafica, por lo tanto todo cambio que se realice sobre ellos en esta ventana será reflejado en su copia en forma de panel. La opción **Agregar Texto** permite agregar una línea de texto, con las mismas características que se explicaron en Solapa Herramientas, Botón Texto. **Agregar series marcadas para mover**, permite agregar las series que fueron marcadas para copiar en la solapa Series.

La Versión Estudiantil de InfoStat no tiene acceso las opciones para guardar gráficos. Además, los gráficos generados con la Versión Estudiantil aparecerán con una marca de agua que hace referencia a la versión.

Suscripción y copia de formatos gráficos

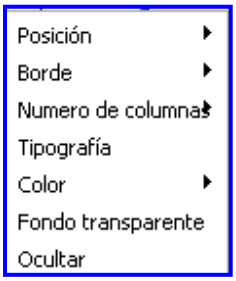
InfoStat clasifica a sus gráficos como *Editores* o *Subscriptores* de formato o de formato *Libre*. El *Editor* de formato es un gráfico que van a ser usado como modelo para copiar sus atributos. Los atributos y los aspectos de esos atributos que se deseen transferir, se especifican utilizando el submenú **Preferencias de copia y suscripción**, al que se puede acceder activando el botón derecho del ratón sobre el gráfico *Editor*. Este submenú activa una ventana de diálogo: **Gráficos: preferencias de suscripción** donde se encontrarán todas las opciones disponibles (solapa **Opciones**). La opción copiar atributos de la serie por nombre o por orden, tiene por objeto asignar características de las series según coincidan por el orden en que se grafican o por su nombre.

Los gráficos *Suscriptores* "apuntan" al Editor del cual copian sus propiedades. Este vínculo es dinámico por lo que cada vez que el *Editor* es modificado, estas características se reflejan en el *Suscriptor*. Para que un gráfico se transforme en suscriptor, primero hay que activarlo haciendo *click* en la solapa numerada correspondiente, que se encuentra en la parte inferior de la ventana. Luego debe seleccionarse el gráfico *Editor* desde el menú que se despliega al activar el botón derecho del ratón **Suscribir al formato del gráfico...**. Todo gráfico puede ser *Editor* excepto los que ya son *Subscriptores*. Existe una alternativa sencilla para declarar simultáneamente varios *Subscriptores* de un mismo *Editor*. Para acceder a ella el *Editor* debe estar activo y se debe desplegar el menú activando el botón derecho del ratón; después se elige la opción **Preferencias de copia y suscripción**. En la solapa **Suscriptores** se pueden seleccionar los *Suscriptores* al *Editor* elegido.

Un gráfico puede dejar de ser *Suscriptor* en cualquier momento, en ese caso sus atributos quedarán desvinculados de los del gráfico *Editor*. Esto se consigue con el comando **Cancelar suscripción de formato** del botón derecho del ratón. Para que el mecanismo de suscripción funcione adecuadamente con el eje X categórico, los cambios en los nombres de las categorías y en la ubicación de las mismas en el eje del editor, se deben realizar posteriormente a la suscripción. Asimismo, es posible copiar los atributos de un gráfico, sin necesidad de suscribirse, simplemente usando el submenú **Copiar formato del gráfico..**.

Observación: Si el eje X representa una variable categórica y se pretende cambiar la posición de las distintas categorías en el gráfico Editor para que estos cambios se reflejen en los gráficos suscriptores, entonces activar las celdas de "copiar mínimo y máximo" del eje X.

Leyendas



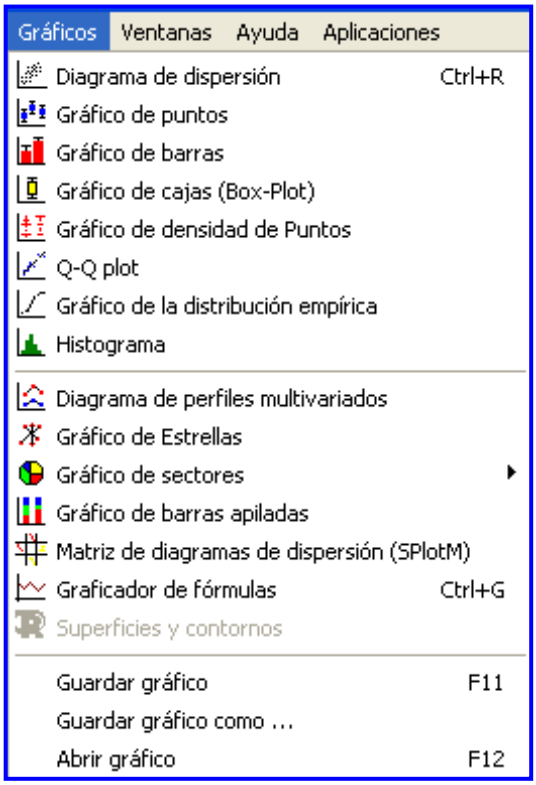
La leyenda de un gráfico es un objeto con su propio menú de opciones, que se activa tocando la leyenda con el ratón y apretando el botón derecho. Las opciones de la leyenda son: **Posición** (**Derecha**, **Izquierda**, **Arriba**, **Abajo** y **Libre**), **Borde** (**Visible**, **Invisible**) alterna entre presencia y ausencia de recuadro respectivamente, **Número de columnas**

permite modificar la forma de presentación de las leyendas, **Tipografía** (abre una ventana de dialogo para cambiar tipografía), **Color** (permite elegir un color de fondo para la leyenda), **Fondo transparente** pone la leyenda sin color y **Ocultar** para no ver la leyenda. Para cambiar los títulos de la leyenda, el orden o sus íconos, se deben editar esas características sobre las series gráficas en la ventana de **Herramientas gráficas**, solapa **Series**.

Líneas de texto

La introducción de líneas de texto en los gráficos se describió anteriormente en **Botón Texto**. Los rótulos, el título del gráfico y las leyendas de los ejes, son todos objetos editables con opciones que se presentan en el botón derecho del ratón una vez que están seleccionados. (Se puede **Copiar**, **Pegar** u **Ocultar** el texto, cambiar la **Tipografía**, la **Orientación**, el **Color**, el **Fondo**, o **Editar** el texto seleccionado. Las leyendas de los ejes y el título del gráfico tienen posiciones por defecto que se anulan cuando estos rótulos son movidos. Para recuperar esas posiciones existe la opción **Reset** en el submenú asociado. Los rótulos numéricos de una escala no pueden modificarse ni moverse. Para modificar la tipografía por defecto de las líneas de texto ver **Botón Texto**.

Además de las opciones propias de la ventana **Gráficos** existen un conjunto de acciones que se pueden realizar en forma directa sobre los gráficos creados. Un Gráfico de InfoStat es una colección de objetos, cada uno de ellos interrelacionados pero con posibilidades de ser modificados independientemente.



Algunos de estos objetos, como Eje X y Eje Y tienen sus opciones en la ventana Herramientas Gráficas, otros pueden modificarse directamente desde el gráfico. Por ejemplo, los puntos que forman un diagrama de dispersión, pueden modificarse en forma general en la solapa **Series** de Herramientas gráficas, pero se pueden modificar en forma individual desde el gráfico. Si uno toca estos puntos con el botón izquierdo del ratón, se colocara a su lado el número del caso en el archivo de origen. Si uno toca un punto con el botón derecho del ratón, aparecerá **Color de ítem gráfico** y esto despliega una barra de colores para cambiar el color de ese punto en particular.

InfoStat puede realizar los siguientes gráficos. A ellos se accede a través del menú **Gráficos** de la ventana principal de InfoStat, estos son: **Diagrama de dispersión**, **Gráfico de puntos**, **Gráfico de**

barras, **Gráfico de cajas (box-plot)**, **Gráfico de densidad de puntos**, **Q-Q plot**, **Gráfico de la distribución empírica**, **Histograma**, **Diagrama de perfiles multivariados**, **Gráfico de estrellas**, **Gráfico de sectores**, **Gráfico de barras apiladas** y **Matriz de diagramas de dispersión**. También dispone de un **Graficador de fórmulas**.

Al igual que en el menú Estadísticas, el menú Gráficos de InfoStat presenta dos ventanas de diálogo. La primera (selector de variables) sirve para establecer las variables que serán utilizadas para construir el gráfico, para definir particiones o especificar algunos atributos como tamaño y rótulos de los elementos gráficos. La segunda (ventana de opciones) tiene por objeto ajustar diversas características propias de los distintos tipos gráficos, e indicar si los gráficos que se producen por particiones estarán en el mismo gráfico o en gráficos separados. Esta ventana puede no aparecer si el gráfico no requiere especificación de opciones.

A continuación se presentan los distintos tipos gráficos. Debido a la gran variedad de opciones y combinaciones que InfoStat ofrece para realizar presentaciones gráficas, la descripción de los distintos tipos se hará presentando inicialmente las características generales del gráfico y luego con ejemplificaciones se mostrarán distintas opciones y detalles para cambiar su aspecto.

Diagrama de dispersión

El diagrama de dispersión es el típico gráfico que muestra un conjunto de puntos ordenados en el plano por sus coordenadas X e Y. Se utiliza cuando se quiere visualizar la variación conjunta de dos variables cuantitativas. Sólo se puede representar una variable en el Eje X pero pueden graficarse simultáneamente varias variables en el Eje Y.

Los siguientes diagramas de dispersión muestran la relación entre porcentaje de germinación (PG) y porcentaje de plántulas normales (PN) que se encuentran en el archivo *Atriplex.*

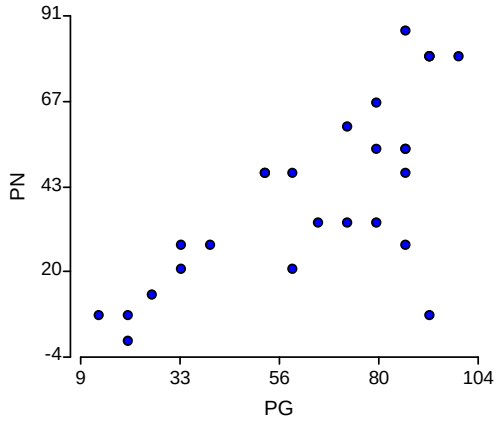


Figura 43: Diagrama de dispersión de plántulas normales (PN) versus porcentaje de germinación (PG). Archivo Atriplex.

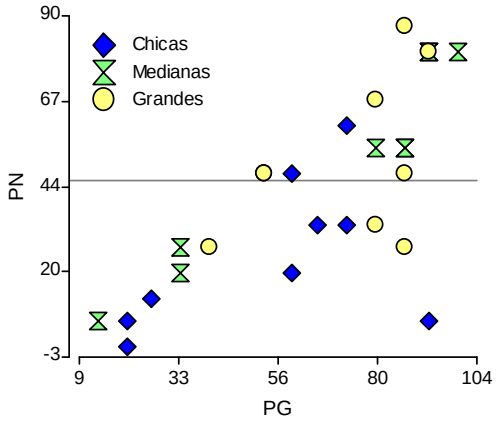


Figura 44: Diagrama de dispersión de plántulas normales (PN) versus porcentaje de germinación (PG). Archivo Atriplex.

La Figura 43 muestra un diagrama de dispersión elemental. Esta figura se obtuvo asignando en la ventana de selección de variables "PN" al eje Y y "PG" al eje X.

La Figura 44 también se obtuvo haciendo un diagrama de dispersión de "PN" versus "PG", pero utilizando el "tamaño" de las semillas como criterio de partición. Los gráficos que se generan para los distintos tamaños de semilla se dispusieron en la misma figura indicando, en la ventana de diálogo, que las particiones se incluyeran en el mismo gráfico. Para identificar los diferentes tamaños de semilla, se utilizaron puntos de distinto color (seleccionar la **Serie** y con el botón derecho seleccionar **Color***)* y distintos símbolos (seleccionar la serie y con el botón derecho seleccionar **Símbolo**). El **Tamaño** de los elementos gráficos se incrementó a 10 puntos. Se incluyó un punto de corte sobre el eje Y correspondiente 45 plántulas normales ("PN"). También se agregó una leyenda (sobre el gráfico apretar botón derecho y seleccionar **Mostrar leyenda**) para asociar los símbolos y colores con los tamaños de semillas (particiones). Esta leyenda fue colocada en **Posición> Libre**, y de esa forma se la pudo llevar dentro del grafico. Además, sobre la leyenda presionando el botón derecho del ratón en el menú que aparece se eligió **Borde**, **Invisible**.

Cuando se tiene un diagrama de dispersión como el que se presentó en el gráfico de la Figura 43, se puede ajustar una recta o una curva que sintetice la relación que se observa entre las variables. Seleccionando una (o varias) de las series y apretando el botón derecho del ratón se puede pedir que InfoStat aplique un suavizado. Esta opción genera una nueva **serie suavizada** por cada una de las series seleccionadas (Figura 45). Por defecto InfoStat aplica como suavizado la regresión localmente ponderada (LOWESS) (Cleveland, 1979). Cuando una serie suavizada se selecciona, aparecen opciones en la ventana de herramientas gráficas que permiten cambiar el suavizado o especificar distintos parámetros para los mismos y actualizar la figura de acuerdo a ellos. En la Figura 46 se muestra un suavizado LOWESS que ha sido cambiado por el ajuste de una recta (polinomio de orden 1).

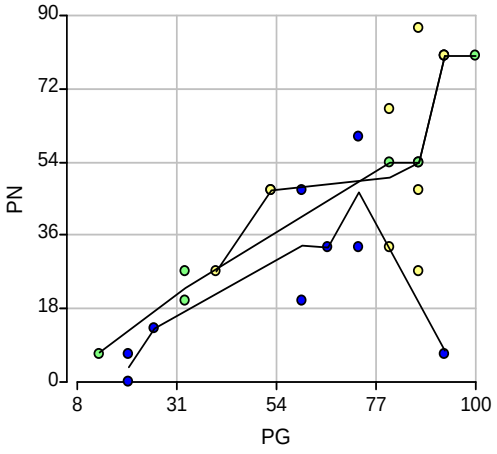


Figura 45: Diagramas de dispersión para plántulas normales versus germinación (%) en semillas de distinto tamaño. Suavizado: regresión localmente ponderada (LOWESS). Archivo Atriplex.

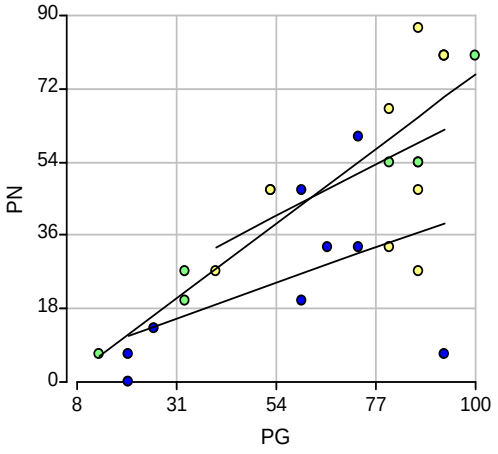


Figura 46: Diagramas de dispersión para plántulas normales versus germinación (%) en semillas de distinto tamaño. Suavizado: polinomio de orden 1. Archivo Atriplex.

Gráfico de Puntos

Este tipo gráfico es similar al diagrama de dispersión y tiene por objeto mostrar una medida resumen de un conjunto de datos y no valores individuales. Lo más común es tratar de representar valores medios de una variable "Y" en relación a una variable cuantitativa o categórica "X". Si se asigna al eje X una variable cuantitativa, InfoStat da la opción de tratarla como categórica, en cuyo caso cada valor diferente es considerado una nueva categoría (opción por defecto).

Estos gráficos, al igual que los diagramas de barra, pueden tener asociados segmentos de recta que representan medidas de variabilidad (por defecto, el error estándar de la media). A continuación se presenta un diagrama de puntos para el diámetro del cuerpo de un nematodo ("Diamcpo") que crece a distintas temperaturas ("TEMP"). Para ello se utilizó el archivo *Hembras,* (gentileza Dr. M. Doucet, Facultad de Ciencias Agropecuarias-U.N.C), ingresando la variable Diamcpo como **Variable a graficar** y la variable TEMP como **Criterio de clasificación**.

Los puntos que representan los valores medios pueden "conectarse", para ello seleccione la serie que quiere conectar y luego, con el botón derecho del ratón activar **Conectores**>**Visibles**. Los conectores sirven para dar una idea del nivel de la variable en puntos intermedios de temperatura, como se muestra en las siguientes figuras:

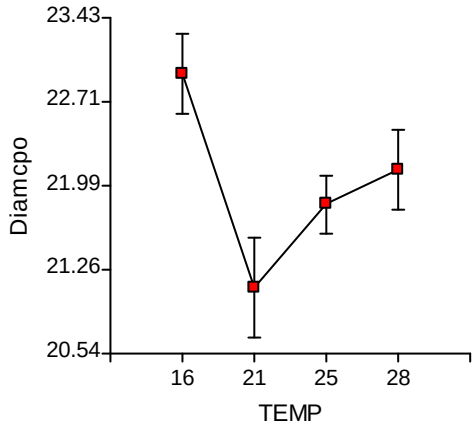


Figura 47: Diagrama de puntos, con "conectores visibles", para el diámetro del cuerpo de nemátodos. Archivo Hembras.

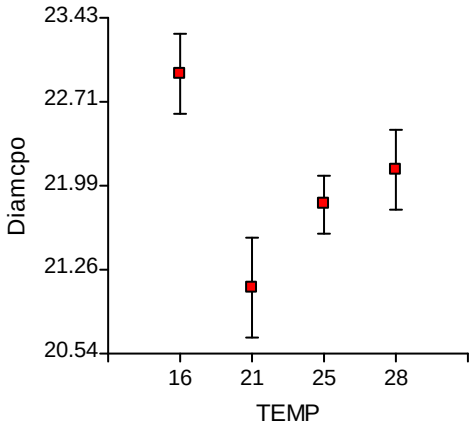


Figura 48: Diagrama de puntos, sin conectores, para el diámetro del cuerpo de nemátodos. Archivo Hembras

A estos gráficos de puntos se les pueden agregar bandas de confianza o de predicción, ya sean paramétricas o no paramétricas. Estas bandas se obtienen uniendo los puntos extremos de las barras que representan los errores. Este efecto se logra haciendo visibles los **Envolventes**. Para ello seleccionar la serie que se desea envolver y luego con el botón derecho del ratón seleccionar **Envolvente>Visible** (Figura 49). Como los puntos y los errores se pueden hacer invisibles (con la serie seleccionada buscar en el menú del botón derecho **Cuerpo> Vacío** y **Sin borde** y **Errores >Invisibles>Ambos**) para conseguir una banda de confianza (o predicción), limpia de otros elementos gráficos (Figura 50).

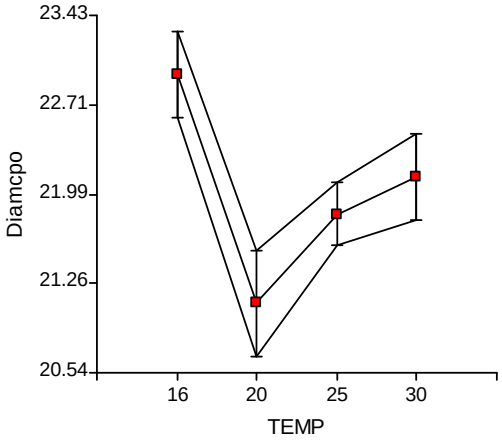


Figura 49: Diagrama de puntos, con "envolventes visibles", para el diámetro del cuerpo de nemátodos. Archivo Hembras.

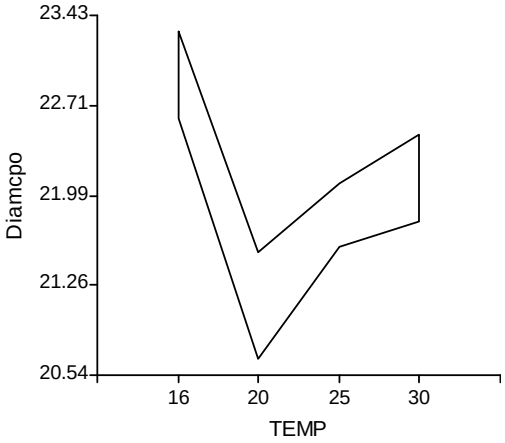


Figura 50: Diagrama de puntos, con "bandas de confianza", para el diámetro del cuerpo de nemátodos. Archivo Hembras.

Gráfico de barras

Este diagrama representa valores medios (opcionalmente medianas, frecuencias absolutas, relativas, mínimos o máximos) de una o más variables en relación a una o más variables de clasificación. Se puede agregar a la representación de los valores medios, una medida de variabilidad que puede ser el error estándar muestral de la media (valor por defecto), la desviación estándar muestral, un intervalo de confianza para la media para un nivel de confianza dado, un intervalo de predicción paramétrico o no paramétrico o una medida de variación arbitraria. Las figuras que siguen representan valores medios del porcentaje de germinación de semillas de distinto tamaño, usando los datos del archivo *Atriplex*.

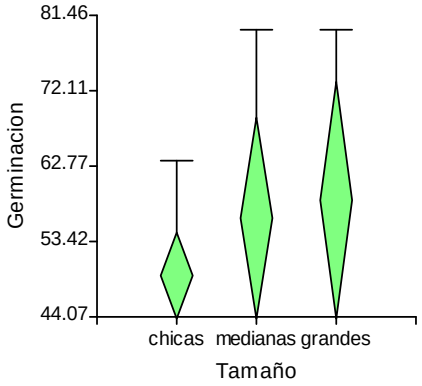


Figura 51: Gráficos de barras, con "símbolo rombo", para el porcentaje de germinación de semillas de distinto tamaño. Archivo Atriplex.

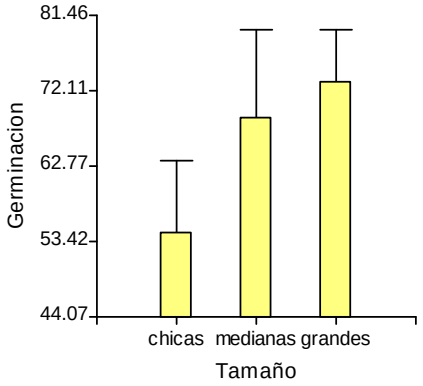


Figura 52: Gráficos de barras, con "símbolo cuadrado", para el porcentaje de germinación de semillas de distinto tamaño. Archivo Atriplex.

Si se especifican uno o más criterios de partición, InfoStat da la opción de generar un gráfico para cada partición o poner todas las particiones en el mismo gráfico. El siguiente es un ejemplo que muestra para los datos del archivo *Atriplex,* el porcentaje de germinación de semillas chicas, medianas y grandes según el color del tegumento de las semillas. Para ello se realizó un diagrama de barras del porcentaje de germinación (seleccionado como Variable Y), en relación al tamaño de las semillas que se declaró como Variable X. El color del tegumento se especificó como partición (Figura 53).

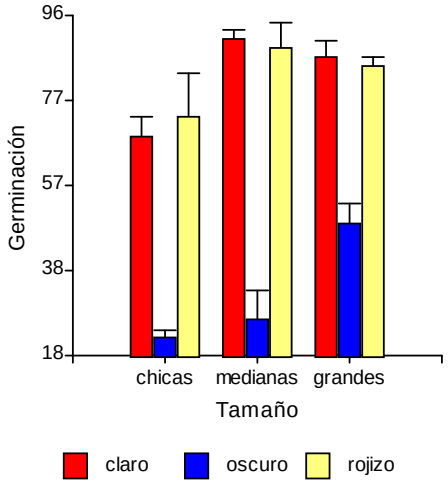


Figura 53: Gráficos de barras para germinación en relación al tamaño de la semilla y partición por color. Archivo Atriplex

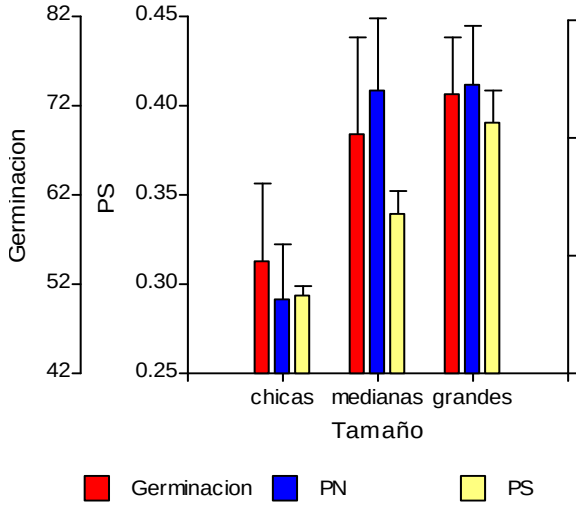


Figura 54: Gráficos de barras para germinación, peso seco y plántulas normales de semillas, según su tamaño. Archivo Atriplex

Así como una variable puede representarse en varias condiciones y según diversos criterios de partición en un mismo gráfico, es posible visualizar simultáneamente varias variables en un mismo gráfico. Como las variables pueden medirse en escalas no compatibles, InfoStat permite agregar tantos ejes como series tenga el gráfico. La Figura 54 representa los valores medios del porcentaje de germinación, porcentaje de plántulas normales y peso seco de las plántulas obtenidas de semillas de distinto tamaño.

En la solapa **EjeY** se encuentra el campo **Base** en el que se puede escribir un valor de referencia para lograr gráficos de barras que se proyecten por encima y por debajo de este valor. Al poner un valor de base, todos los valores de respuesta para cada valor en el eje X que superen el valor base propuesto, quedarán graficados por encima de ese valor, mientras que los que no lo superen, serán graficados por debajo representando la diferencia con respecto al valor base. El campo de edición de la base está protegido y cualquier modificación no tendrá efecto mientras la protección esté activada. Para anular la protección, se debe desactivar la casilla que se encuentra a la izquierda del campo, haciendo *click* con el cursor posicionado sobre ésta, luego escribir el valor para la base. Usando el archivo *Atriplex*, la Figura 55 muestra un gráfico de barras para plántulas normales en relación al color del episperma de la semilla, con base igual al valor mínimo, en este caso 10, y la Figura 56 uno similar pero con base igual a 40. En la Figura 55 se marcaron en el eje Y tres líneas de corte: 62, 43 y 22, para representar el número de plántulas normales

promedio para cada color de episperma. Obsérvese que en la Figura 56 se grafica la diferencia con respecto al valor 40.

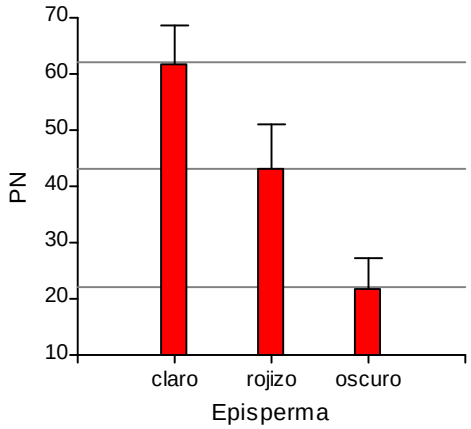


Figura 55: Gráfico de barras para plántulas normales en relación al color del episperma de la semilla, con base igual al valor mínimo. Archivo Atriplex.

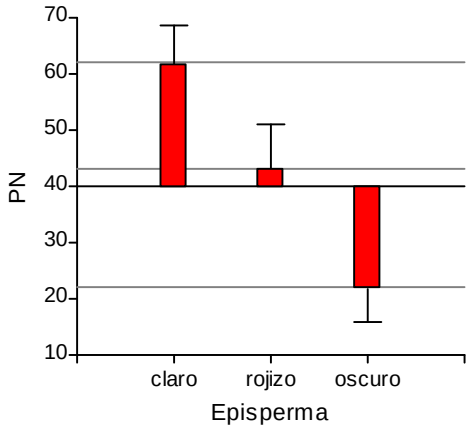


Figura 56: Gráfico de barras para plántulas normales en relación al color del episperma de la semilla, con base diferente al valor mínimo. Archivo Atriplex.

Gráfico de cajas (box-plot)

Los diagramas de barras y de puntos resumen la información muestral en un punto. Eventualmente pueden incluir una medida de error o variabilidad. Sin embargo es difícil, a partir de ellos, visualizar la forma de la distribución de frecuencias de cada grupo de observaciones. El gráfico de cajas *(box-plot)*, tiene por objeto reflejar mejor la forma de estas distribuciones dando en un mismo elemento gráfico información acerca de la mediana, la media, los cuantiles 0.05, 0.25, 0.75 y 0.95 y mostrando la presencia, si los hubiere, de valores extremos. La especificación de las variables en el selector de variables de este tipo de gráfico es idéntica que para el diagrama de puntos.

En la Figura 57 se representa el diagrama de cajas para la variable diámetro del cuerpo ("Diamcpo") de nemátodos que crecen a distintas temperaturas (archivo *Hembras*). Para las temperaturas de 21 y 28 °C se indican sobre el gráfico qué representan cada uno de los detalles del diagrama de cajas. Los rótulos fueron agregados con la herramienta **Texto** de la ventana de **Herramientas gráficas**.

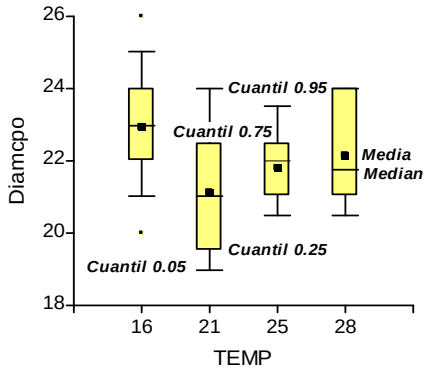


Figura 57: Gráficos de cajas (box-plot), para diámetro del cuerpo ("Diamcpo") de nematodos a distintas temperaturas. Archivo Hembras.

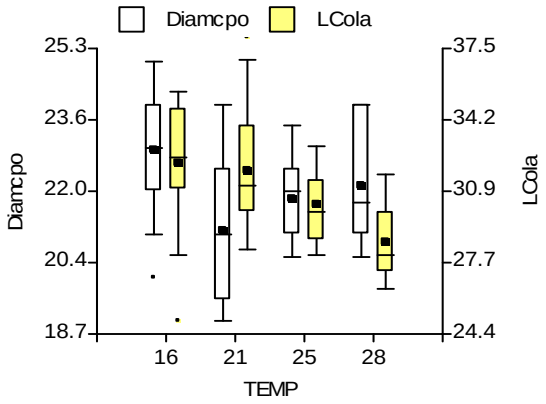


Figura 58: Gráficos de cajas (box-plot), para diámetro del cuerpo ("Diamcpo") y longitud de la cola ("Lcola") de nematodos a distintas temperaturas. Archivo Hembras.

Al igual que en el diagrama de barras, es posible incluir en un mismo gráfico varias variables, cada una (si fuere necesario) con su propio eje de coordenadas Y, y/o varias particiones. En la Figura 58 se graficó simultáneamente el diámetro del cuerpo ("Diamcpo") y el largo de la cola ("Lcola") de los nemátodos de este ejemplo.

Cuando se selecciona una serie de diagramas de cajas, la ventana **Herramientas gráficas**, solapa **Series** presenta un conjunto de opciones que permiten incluir un símbolo que representa la media muestral (opción activada por defecto) y especificar criterios para visualizar valores extremos.

Gráfico de densidad de puntos

Aunque los diagramas de cajas dan importante información sobre la forma de la distribución de frecuencias, a veces es muy útil ver de manera directa donde están los casos efectivamente observados, especialmente si su número es pequeño. La especificación de las variables para este gráfico es idéntica que para el diagrama de puntos. La Figura 59 presenta un diagrama de densidad de puntos para el porcentaje de germinación de semillas de distinto tamaño (archivo *Atriplex*). En este gráfico se han incluido puntos de corte en los porcentajes de germinación del 10, 50 y 90%, ingresando dichos valores en forma de lista en la ventana **Herramientas gráficas, solapa Eje Y**, panel **Líneas de corte**. En el punto de corte del 50%, se utilizó un mayor espesor. Para ello seleccionar el punto de corte con el ratón y en el menú que se despliega con el botón derecho seleccionar **Espesor**. Si se desea dar color, seleccionar **Color** en el mismo menú.

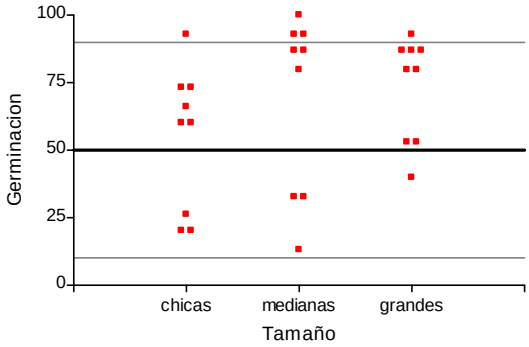


Figura 59: Gráfico de densidad de puntos para la germinación de semillas de distinto tamaño. Archivo Atriplex.

Q-Q plot

Estos gráficos se utilizan para evaluar el grado de ajuste de un conjunto de observaciones a una distribución teórica. Aunque no representan pruebas formales de ajuste, la experiencia ha mostrado que son efectivos para detectar faltas de ajuste que muchas veces las pruebas formales son incapaces de detectar.

Los parámetros de la distribución teórica seleccionada son estimados a partir de la muestra. La ventana de opciones de **QQ-Plot** permite seleccionar entre varias distribuciones teóricas: Normal, Chi cuadrado, Exponencial, Weibull, Gumbel y Beta. Si se desea, se puede representar en el grafico la recta Y=X, activando el campo **Mostrar recta Y=X**.

El siguiente es un ejemplo de Q-Q plot normal aplicado a una muestra de 50 observaciones provenientes de una distribución normal (archivo *Qqplot*). Los datos corresponden a las longitudes de pétalos de 50 flores. En este gráfico se ha agregado la recta Y=X.

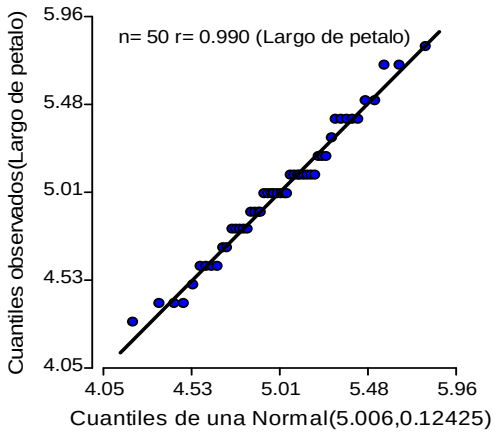


Figura 60: Q-Q plot normal para 50 observaciones de la variable largo de pétalo. Archivo Qqplot.

En el gráfico Q-Q plot, se presentan en la parte superior el tamaño muestral (*n*) y el coeficiente de correlación lineal *r* de la correlación entre los cuantiles observados versus los cuantiles de la distribución teórica seleccionada. Sobre el rótulo del eje X se muestran los parámetros de la distribución teórica estimados a partir de la muestra por máxima verosimilitud.

Gráfico de la distribución empírica

Aunque el polígono de frecuencias relativas acumuladas (disponible como opción del submenú **Histograma**) puede utilizarse para visualizar la forma de la función de distribución empírica de un conjunto de observaciones, la técnica requiere contar con un número usualmente grande de datos para que represente una buena aproximación de la verdadera distribución.

Existen muchas técnicas de construcción de gráficos de distribución empírica de una variable. El gráfico que genera InfoStat se basa en calcular la distribución según el siguiente algoritmo: Sean $x^{(1)}$, $x^{(2)}$, $x^{(n)}$ las observaciones de una muestra de tamaño *n* ordenadas de menor a mayor. La función de distribución empírica evaluada en la observación $x^{(i)}$ se calcula como $F(x^{(i)})=(i- 0.375)/(n+0.25)$.

Este gráfico presenta los valores observados de la variable en el eje X y la función de distribución empírica evaluada en cada uno de los puntos observados en el eje Y.

A continuación se presentan dos ejemplos de gráficos de la función de distribución empírica. La Figura 61 muestra el gráfico de la distribución empírica de la longitud del pétalo en la muestra conjunta que incluye tres especies de *Iris* (*setosa, versicolor y virginica*), archivo *Iris*). El gráfico tiene la grilla activada para ejemplificación. La longitud del pétalo es una de las variables que mejor diferencian estas especies. El gráfico de la función de distribución empírica para la muestra conjunta evidencia una fuerte anomalía, con respecto a una función de distribución normal, indicando la posible mezcla de distribuciones, ya que se grafican datos posiblemente provenientes de tres distribuciones normales distintas, cada una asociada a una especie.

La Figura 62 corresponde al perímetro de cabezas de ajo blanco de la campaña 1998 del archivo *AjoBlanc* (gentileza Dra. V. Conci, IFFIVE-INTA). En este último gráfico se han agregado dos líneas de corte que permiten identificar fácilmente el percentil 50% que se corresponde con el valor 15. Dado el gran número de casos, el tamaño de los puntos se redujo a 2. Eventualmente los puntos podrían hacerse desaparecer reduciendo su tamaño a 0 y perfilar la forma de la distribución poniendo a los conectores visibles. Opcionalmente se podría pedir a InfoStat que genere una serie suavizada para estos gráficos.

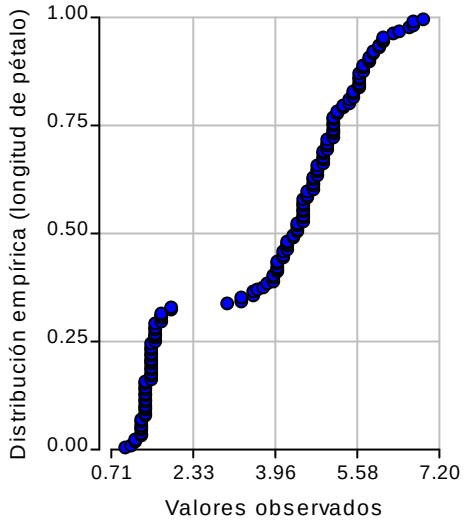


Figura 61: Gráfico de la distribución empírica de la longitud de pétalos. Archivo Iris.

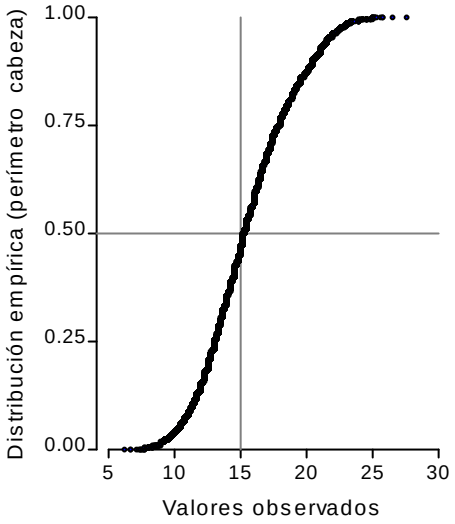
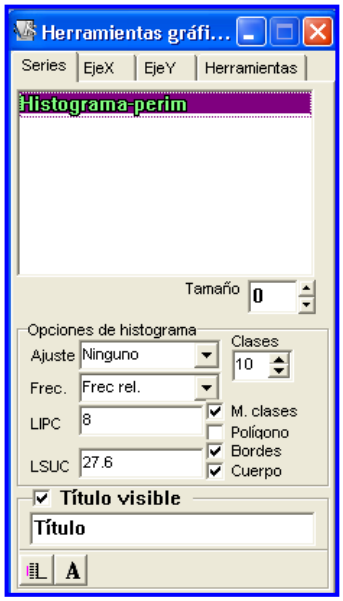


Figura 62: Gráfico de la distribución empírica del perímetro de la cabeza de ajo blanco. Archivo AjoBlanc.

Histograma

InfoStat permite construir histogramas de frecuencias cuando se tienen suficientes observaciones. Estos histogramas pueden ser utilizados para aproximar la distribución teórica subyacente. Experiencias prácticas señalan que una amplia gama de distribuciones pueden aproximarse bien desde una distribución empírica construida a partir de 50 o más observaciones. Al construir un Histograma, la ventana **Herramientas Gráficas** muestra un diálogo que permite modificar los atributos del histograma obtenido. En la solapa **Series** de esta ventana, hay un menú de opciones de histograma que permite: cambiar el número de clases (**Clases**) que por defecto son calculadas como $Log_2(n+1)$; realizar ajustes (**Ajuste**) a una distribución Normal, Chi Cuadrado, Exponencial, Weibull, Gumbel, Beta y Gamma con parámetros iguales a las estimaciones muestrales obtenidos por máxima verosimilitud. En caso de no desear un ajuste en el campo debe figurar la palabra **Ninguno**. Además la ventana permite elegir la frecuencia representada en el histograma (**Frec**.). Las frecuencias a graficar pueden ser: frecuencia relativa (**Frec. rel**) que es la opción por defecto, frecuencia absoluta (**Frec. abs**.), frecuencia absoluta acumulada (**Frec. abs. acum**.) y frecuencia relativa acumulada (**Frec. rel. acum**.). El campo **Borde** permite eliminar los contornos de las barras que forman el histograma. Se puede construir el



polígono de frecuencia activando el campo **Polígono**. El campo **Cuerpo** permite eliminar el histograma de fondo a partir del cual se dibujó el polígono. Los campos **LIPC** y **LSUC** permiten ingresar los límites inferior y superior para la primera y última clase respectivamente. Para lograr que los "ticks" se correspondan con las marcas de clase de cada intervalo activar **M. clases** en la solapa **Series**.

A continuación se presentan los histogramas de frecuencias relativas (Figura 63) y relativas acumulada (Figura 64) para la variable perímetro de cabezas de ajo blanco de un ensayo de sanidad vegetal (archivo *AjoBlanc*). En la Figura 64 se han agregado puntos de corte en el eje Y para el valor Y=0.50 y sobre el eje X para el valor X=17. En ambos gráficos se ha utilizado la técnica de doble fila para la escala del eje X a fin de que los rótulos de la escala no se superpongan.

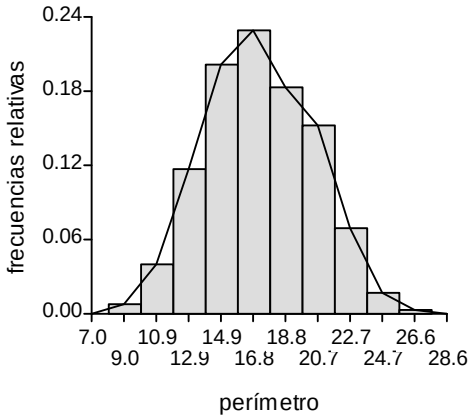


Figura 63: Histograma y polígono de frecuencias relativas para el perímetro de cabezas de ajo blanco (archivo AjoBlanc).

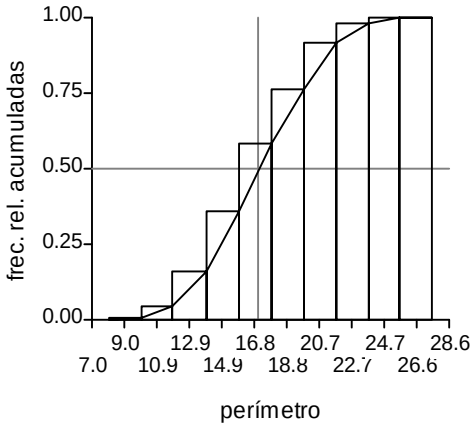


Figura 64: Histograma y polígono de frecuencias relativas acumuladas para el perímetro de cabezas de ajo blanco (archivo AjoBlanc).

Diagrama de perfiles multivariados

Cuando se tienen medidas de una variable repetidas en el tiempo o de varias variables medidas en un mismo individuo o unidad experimental, puede ser de interés visualizar como es la forma de los perfiles respuesta.

InfoStat puede graficar para cada variable, o para una variable en distintos tiempos, un punto o una barra que representa su valor medio, un box-plot que resume la forma de su distribución o un gráfico de densidad de puntos. Esta técnica de graficación requiere que las distintas variables o medidas en el tiempo cuyo perfil se quiere graficar, estén en columnas diferentes de la tabla de datos como se muestra en la siguiente ventana, que corresponde a los datos del archivo *Procedencias*.

1. Procedencias

caso	PROC	D030	D060	D090	D120	D197	D324	D434	D498
1	1	19.13	39.40	50.67	50.42	53.00	46.00	72.83	82.00
2	1	16.15	34.35	43.22	44.42	45.83	40.25	76.50	83.83
3	1	16.88	38.88	48.67	50.78	57.92	45.58	83.50	97.33
4	1	19.07	41.02	51.50	52.43	54.75	50.25	81.00	84.33
5	1	16.70	39.38	50.77	44.60	56.17	39.25	73.83	84.67

Real | registros: 138

Ejemplo 48: En plantas de algarrobo la altura de las mismas se registró en 9 oportunidades desde la siembra hasta el día 498. Se utilizó como criterio de clasificación la procedencia de las semillas. Hay 23 procedencias y para cada una de ellas hay seis registros. Los datos se encuentran en el archivo Procedencias (gentileza Ing. Agr. G. Verzino, Facultad de Ciencias Agropecuarias, U.N.C.).

Para generar un perfil multivariado, asignar todas las variables del perfil a la lista de variables de la ventana de selección de variables. Se puede especificar un criterio de clasificación si se desea observar más de un perfil sobre un mismo gráfico. En este caso habrá tantos perfiles como grupos diferentes se obtengan del criterio de clasificación. Por defecto InfoStat propone un diagrama de puntos conectados. Si los perfiles que se muestran son el resultado de promediar la respuesta de varias repeticiones, entonces puede tener sentido hacer visibles las barras de error. Para ello seleccione las series correspondientes y con botón derecho escoja **Errores>Visibles>Ambos**. A continuación se presenta un diagrama de perfiles multivariados para la altura de plantas de algarrobo. En la Figura 65 se muestran los perfiles para todas las procedencias y en la Figura 66 se presentan sólo tres 3 procedencias (perfiles) a los cuales se agregaron barras de errores que representan el error estándar de la media.

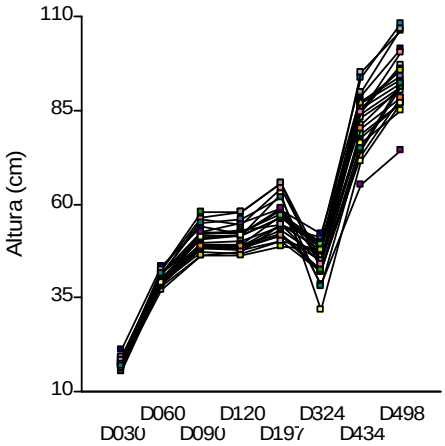


Figura 65: Diagramas de perfiles multivariados para la altura de árboles de algarrobos de 3 procedencias. Archivo Procedencias.

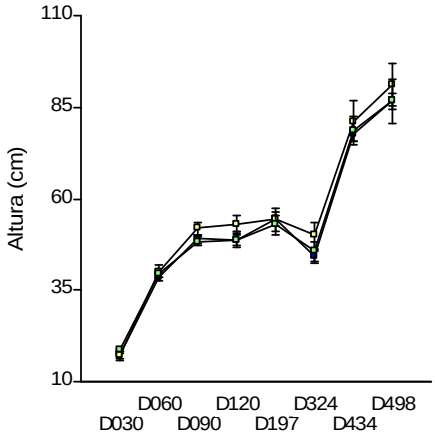


Figura 66: Diagramas de perfiles multivariados para la altura de árboles de algarrobos de 3 procedencias. Archivo Procedencias.

En la siguiente figura, se representa la evolución de altura sin discriminar entre procedencias y utilizando gráficos de caja (box-plot) como elementos gráficos del perfil, los que fueron unidos al hacer visibles los conectores.

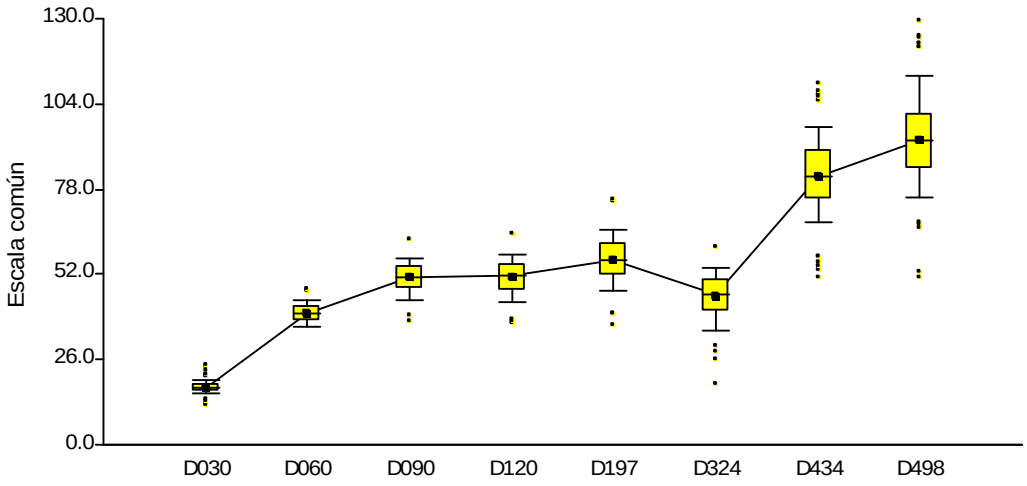


Figura 67: Diagramas de perfiles multivariados, utilizando box-plot, para la variable altura de planta registrada en 9 oportunidades desde la siembra. Archivo Procedencias.

Gráfico de estrellas

Estos gráficos se utilizan para representar de manera concisa y comparativa observaciones multivariadas. Cada variable es representada como un radio de una estrella. La magnitud del radio viene dada por el valor de la variable en la observación representado por la estrella. Si varias observaciones son representadas en una misma estrella, es decir, varias observaciones del archivo tienen el mismo valor para el criterio de clasificación, la longitud del radio es función del valor medio de cada variable. Así, formas de estrellas diferentes indican las variables que marcan mayor diferencia entre las observaciones. Considere el archivo *Iris* que contiene los valores de cuatro variables observadas en flores de 50 individuos de cada una de tres especies del género *Iris*. Aplicando diagramas de estrellas a estas cuatro **Variables** y utilizando como **Criterio de clasificación** a la especie, se obtiene la Figura 68. Para lograr esta figura activar la opción **Estrella con terminaciones.** La Figura 69 es similar a la anterior excepto que tiene además activada la opción **Estrella cerrada**. En ambos casos se activó para las tres series y usando el botón derecho del ratón la opción **Identificadores>Visibles>Rótulos**. Otras variaciones consisten en modificar los colores, los símbolos y la ubicación de las estrellas.

En las figuras puede verse que la variable “PetalWid” es más prominente en la especie número 3, mientras que la variable “SepalWid” lo es para la especie 1. Por otra parte, “PetalLen” y “SepalLen” son comparativamente más pequeñas en la especie 1 que en las especies 2 y 3.

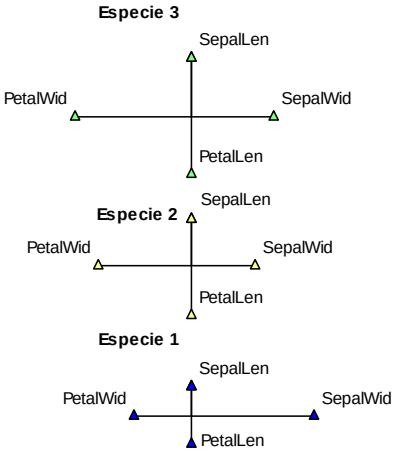


Figura 68: Gráfico de estrellas para las variables longitud y ancho de pétalos y longitud y ancho de sépalos en tres especies del género Iris. Archivo Iris.

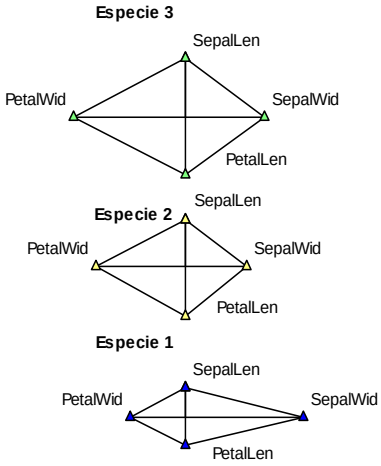


Figura 69: Gráfico de estrellas para las variables longitud y ancho de pétalos y longitud y ancho de sépalos en tres especies del género Iris con la opción de estrella cerradas activa. Archivo Iris.

Si no existiera una variable en el archivo para ser usada como criterio de clasificación, el usuario podrá obtener una "estrella simple", para representar las magnitudes relativas de las variables seleccionadas. Este tipo de estrellas tiene sentido en el caso de variables conmensurables, es decir variables expresadas en una misma escala. En este caso la comparación de las longitudes de los radios permite inferir diferencias entre variables. Cuando se grafica más de una estrella (usando una o más variables como criterios de clasificación) la comparación debe realizarse entre los radios homólogos de las distintas estrellas y no entre los radios de una misma estrella.

Gráfico de Sectores

El gráfico de sectores (tortas) es útil para representar contribuciones porcentuales a un total o la distribución de frecuencias de una variable categórica. Por ejemplo, si los gastos familiares se dividen en alimentación, servicio e impuestos, educación y otros, entonces un gráfico de tortas mostrará la contribución proporcional cada uno de estos rubros en la conformación del gasto total, como puede observarse en la Figura 70.

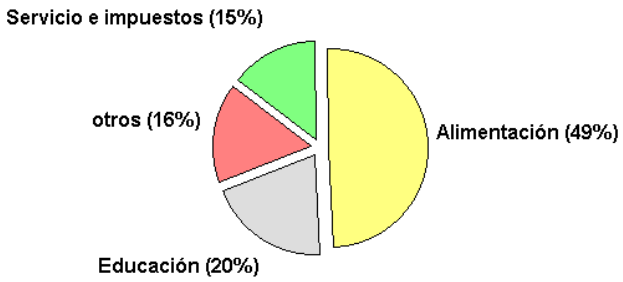
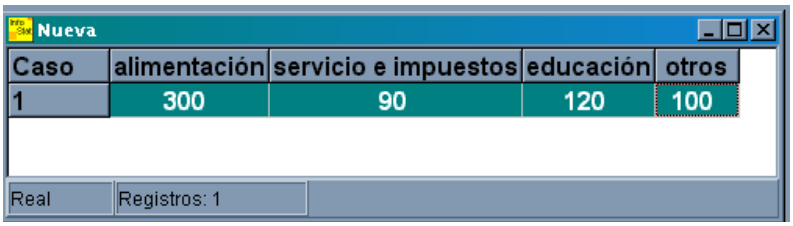


Figura 70: Gráfico de tortas para la distribución de gastos familiares.

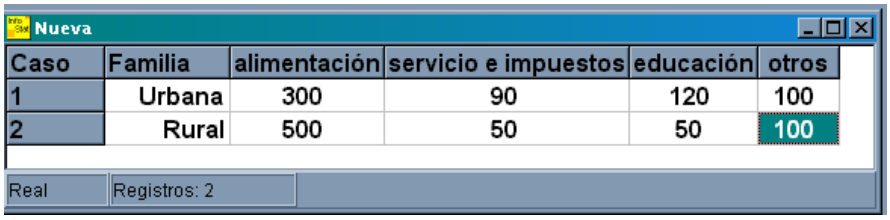
Para construir este gráfico, InfoStat ofrece dos opciones que dependen de la forma en que la información está ordenada dentro de la tabla de datos.

Categorías en columnas: Los datos se presentan disponiendo las distintas contribuciones en las columnas de la tabla, como puede se ilustra en la siguiente ventana. En la ventana de selección de variables para este gráfico las columnas de la tabla, que identifican las contribuciones se deben asignar a la lista **Clases** (sectores de la torta).



Si en la tabla de datos hay más de una fila, InfoStat utilizará los totales por columna para obtener la contribución proporcional de cada ítem. Si por, otra parte, existe un criterio de agrupamiento, es posible poner en un mismo gráfico varias tortas, una por cada criterio de agrupamiento. Asimismo, si para cada uno de ellos se tienen varias filas de datos, entonces el total por columna para cada criterio de agrupamiento será utilizado para construir la torta.

Por ejemplo, si a la tabla anterior se le agrega una columna "Familia" en la que se indica si la familia vive en un ambiente urbano o rural y en la ventana de selección de variables se elige "Familia" como **Criterio de agrupamiento** se obtendrá un gráfico en el que la distribución del gasto aparecerá discriminada según si la familia vive en zona "urbana" o "rural" (archivo *Tortas*).



Las series de un gráfico de tortas representan los distintos sectores. Cuando se grafican varias tortas en un mismo gráfico, aparece una serie adicional llamada *identificación de*

grupos que no se visualiza en el gráfico pero que es la responsable de que aparezcan los subtítulos que identifican a cada torta. Esta serie debe estar al comienzo (por defecto) o al final de la lista de series para que el título de la torta quede ubicado en la parte superior. Si esta serie se hace invisible, entonces los identificadores de cada torta desaparecen.

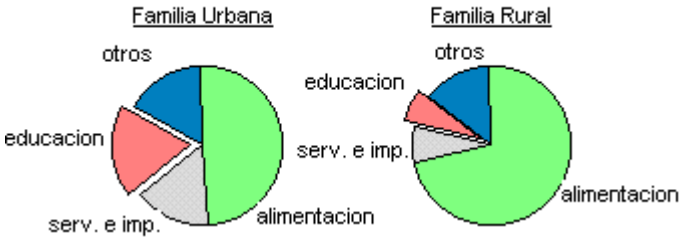


Figura 71: Diagrama de tortas para el ejemplo de ingresos de familias urbanas y rurales. Archivo Tortas.

Seleccionado las series restantes es posible aplicar varias modificaciones al diagrama básico que genera InfoStat. En el panel de opciones, se puede especificar: el **tamaño** general de las tortas (se aplica a todas las series, estén o no seleccionadas), la **separación** entre sectores (en el ejemplo se separó el sector "educación"), si se quiere dar efecto **3D** a uno o más sectores, la posición relativa de los sectores y sus colores. También se puede cambiar selectivamente el contenido de los rótulos asociados a cada sector pudiendo elegir entre 7 formatos diferentes y ninguna rotulación. Editando el nombre de las series se modifican los contenidos de los rótulos.

Cuando el número de tortas en un gráfico es mayor que uno, InfoStat permite que el usuario indique en cuantas columnas se quiere presentar el conjunto de tortas.

La solapa **Orden**, indica el ordenamiento que se tomó al realizar el gráfico. Cuando se generan estos diagramas utilizando un criterio de clasificación, el diagrama muestra varias tortas ordenadas de manera arbitraria. Seleccionando la torta que se quiere mover y aplicando Ctrl.↓ o Ctrl.↑ se pueden ordenar las tortas según la posición deseada.

Las escalas de los ejes X e Y, así como los ejes mismos, están ocultos por defecto ya que en el contexto de los gráficos de torta no tienen ninguna interpretación útil y sólo representan la escala según la cual las tortas son ordenadas. Estas escalas pueden ser modificadas por el usuario si lo cree conveniente, cambiando mínimos y máximos para ajustar detalles de posicionamiento de las tortas. Un hecho importante es que si el usuario cambia el número de columnas o el tamaño de las tortas, InfoStat recalculará automáticamente la escala de estos ejes devolviéndoles sus valores por defecto.

Por otra parte, si se activa la opción **Anillo**, en el ejemplo presentado, se obtiene el siguiente gráfico.

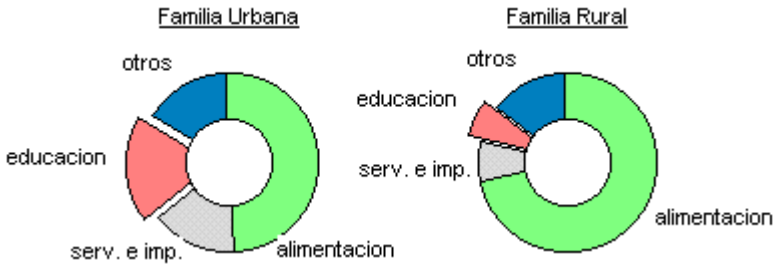


Figura 72: Diagrama de tortas con anillo, para los ingresos de familias urbanas y rurales. Archivo Tortas.

En algunas situaciones el rótulo asociado a uno o varios sectores no cambia cuando se aplica una nueva opción de titulado. Esto puede obedecer a dos causas: la serie correspondiente no está seleccionada o el rótulo ha sido editado. En este último caso, seleccionando el rótulo, apretar el botón derecho de ratón y activar el ítem **Reset**. Con esto, el rótulo volverá a su posición original y podrá ser modificado de acuerdo a la opción de rotulación. El efecto de haber editado o movido un rótulo también afecta su habilidad de "seguir" al sector si este cambia su posición relativa dentro de la torta. Esta situación también se corrige seleccionando el rótulo, apretando el botón derecho del ratón y aplicando el ítem **Reset**. Si se cambia el nombre del sector editando la serie correspondiente en la solapa **Series**, entonces con la opción **Reset** solo ejecutará el cambio a la posición original y las ediciones realizadas *a posteriori* sobre el rótulo, pero manteniendo el nombre de la serie.

Como consejo práctico, se recomienda hacer todas las modificaciones de posicionamiento relativo de sectores, separaciones, ordenamiento en columnas y elegir las opciones de rotulación antes de hacer pequeños ajustes de posicionamiento de los rótulos dentro de la imagen. De esa forma se evitarán efectos indeseados.

La opción **Categorías en filas** para graficar tortas presupone que los datos que identifican a las distintas categorías están en una columna. Lo que InfoStat calculará para armar los distintos sectores de la torta son las frecuencias relativas de cada categoría, ya sea contando el número de cada una de ellas o utilizando sus frecuencias que pueden declararse opcionalmente. También en esta modalidad se pueden obtener tortas para subconjuntos de datos definidos por **Criterios de agrupamiento**. Una vez obtenidas las tortas, todas las opciones discutidas anteriormente son aplicables.

Gráfico de barras apiladas

El gráfico de barras apiladas se aplica cuando se quiere representar comparativamente la contribución que distintas componentes hacen a un total. Por ejemplo, si el peso de una planta se divide en el peso de raíz, tallo y hojas, entonces, la contribución de cada una de estas particiones al peso total se puede representar como los segmentos de una barra cuya altura es 1. Luego, si el peso de las hojas es el que más contribuye al peso total, el segmento asociado al peso de las hojas será el más grande. **Gráfico de barras apiladas** ofrece por defecto la presentación de la contribución de cada partición como proporción del total pero

es posible pedir la representación de los valores absolutos de la contribución de cada partición. En este último caso la altura de las barras es variable.

Las series que conforman los distintos segmentos de las barras apiladas se pueden cambiar de posición relativa y se pueden hacer invisibles. En este caso las proporciones se calculan nuevamente. Eventualmente estos segmentos se pueden destacar haciéndolos más angostos o más anchos cambiando el tamaño de los elementos gráficos. Los conectores visibles reflejan el perfil de variación de cada segmento en la barra.

La Figura 73 presenta la contribución proporcional de las ganancias netas totales por año de la casa matriz y cuatro sucursales de una empresa agropecuaria (archivo *Ganancias*). En este ejemplo se quiere mostrar como las ganancias netas totales por año fueron conformadas con las contribuciones parciales de la casa matriz y las sucursales. La Figura 74 muestra el mismo gráfico pero representando los valores absolutos de esas contribuciones.

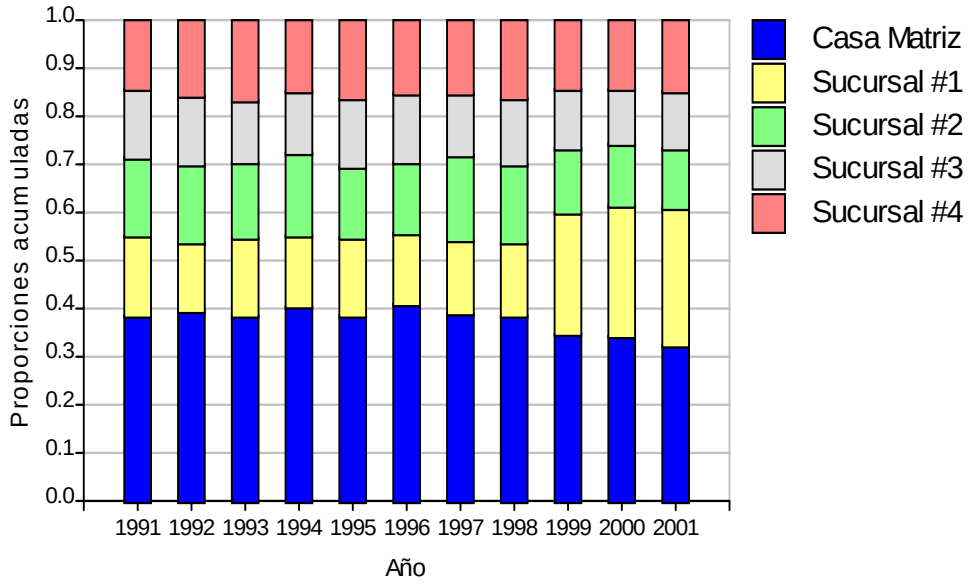


Figura 73: Gráficos de barras apiladas de las contribuciones proporcionales de la casa matriz y cuatro sucursales de un empresa agropecuaria a las ganancias netas discriminadas por año. Archivo Ganancias.

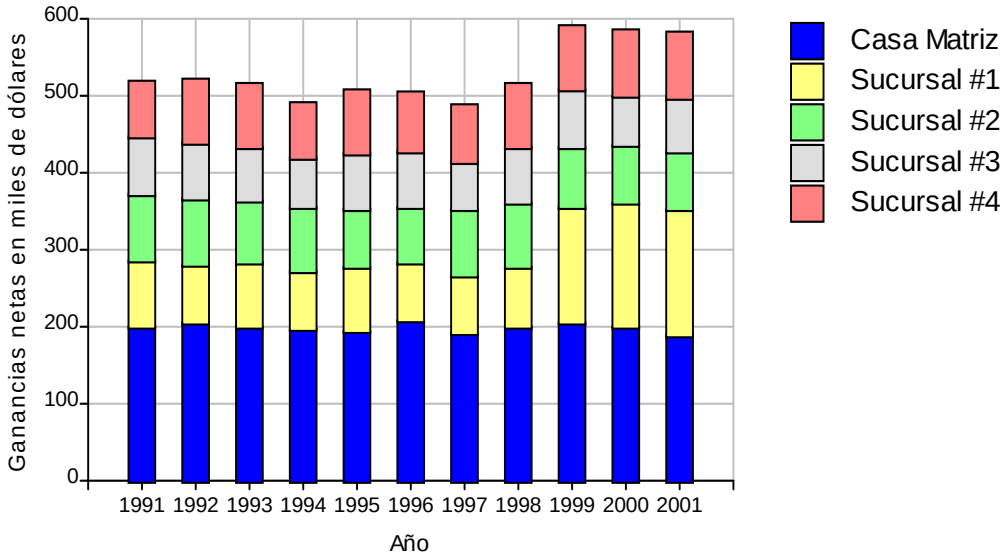


Figura 74: Gráficos de barras apiladas de las contribuciones de la casa matriz y cuatro sucursales de una empresa agropecuaria a las ganancias netas discriminadas por año. Archivo Ganancias.

Matriz de diagramas de dispersión

Estos diagramas permiten producir en un mismo gráfico una matriz de diagramas de dispersión. Esto es útil para visualizar las relaciones entre un conjunto de variables. La Figura 75 muestra en esta forma de representación las relaciones entre la variable "Germinación", plántulas normales (PN) y peso seco de las plántulas (PS) para un ensayo de germinación (archivo *Atriplex*). En la Figura 76 se presenta el mismo gráfico al que se le han agregado suavizados basados en la técnica de regresión localmente ponderada (LOWESS).

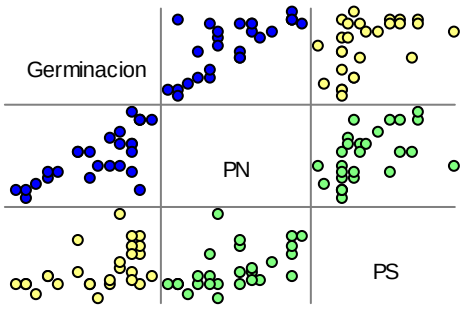


Figura 75: Matriz de diagramas de dispersión para las variables germinación, plántulas normales (PN) y peso seco de las plántulas (PS) en un ensayo de germinación. Archivo Atriplex.

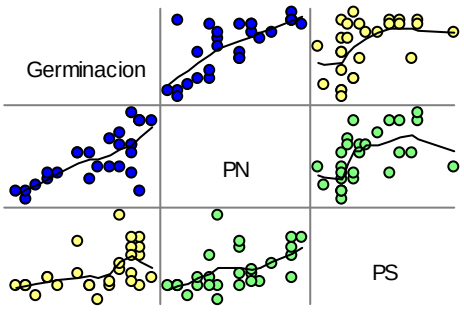
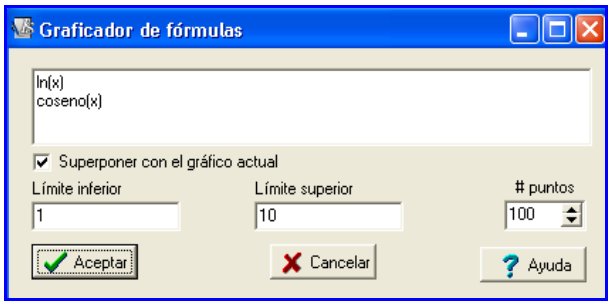


Figura 76: Matriz de diagramas de dispersión para las variables germinación, plántulas normales (PN) y peso seco de las plántulas (PS) en un ensayo de germinación. Las series han sido suavizadas. Archivo Atriplex.

Graficador de funciones

El graficador de funciones es una herramienta gráfica de InfoStat que permite graficar funciones de una sola variable especificadas por el usuario. El usuario puede especificar una a varias funciones simultáneamente. En la ventana de diálogo que se muestra a continuación se han especificado dos funciones de la variable "x", logaritmo natural (ln(x)) y coseno (x). Límite inferior y límite superior son los límites entre los que se graficarán la o las funciones. N puntos es el número de puntos en los que se divide el intervalo de graficación.



De acuerdo a lo especificado en la pantalla anterior el gráfico obtenido es el siguiente:

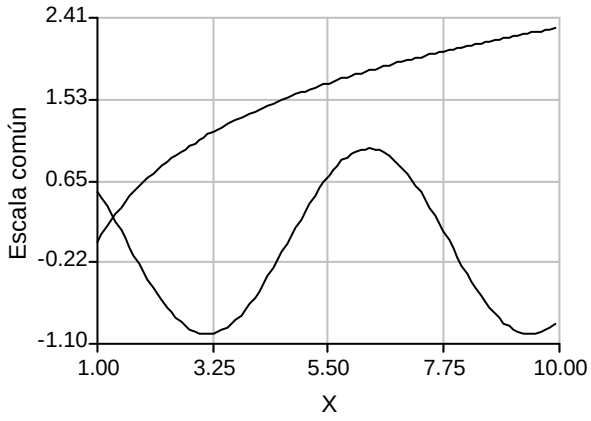


Figura 77: Gráfico obtenido para dos funciones de la variable "x", logaritmo natural (ln(x)) y coseno (x).

Aplicaciones

Control de calidad

En esta versión InfoStat provee al usuario de diagramas de control comúnmente utilizados en el control de la calidad de la producción de bienes y servicios. La versión *full* de la Aplicación Control de Calidad de InfoStat ofrece un menú completo de acciones para realizar control estadístico de calidad. En este manual sólo se documentan las técnicas ofrecidas en esta versión.

La calidad de un producto o servicio se define como su aptitud para el uso demandado por el mercado. Los procesos de producción pueden ser controlados a partir de mediciones de una o más características de calidad. Los parámetros o características de calidad son aquellos atributos o variables del producto que describen su aptitud. Un concepto clave es el de variabilidad o dispersión (diferencias entre los valores de un conjunto de mediciones) de dichas mediciones. La componente aleatoria de cada medición se supone surge del agregado de diversas componentes aleatorias provenientes de distintas fuentes de variación y/o error. La variación total de un conjunto de mediciones se puede descomponer en una suma de mediciones de variación debida a las fuentes que afectan al proceso. Es importante diferenciar entre causas comunes y causas especiales de variación. Las causas comunes pueden ser producidas por numerosos factores que, si bien afectan la distribución de la característica medida, no impiden que el sistema sea predecible. Las fuentes especiales de variación generan una variación no controlada, inestable y no predecible. Las herramientas estadísticas para el control de calidad pretenden reducir la variabilidad de los parámetros de calidad a través del análisis de procesos y la comparación con estándares preestablecidos para proveer información útil en el diseño de acciones tendientes a corregir problemas ocasionados por fuentes de variación especiales. En el marco del control estadístico de calidad se llama *sistema estable de causas fortuitas* a aquel donde las características de calidad muestran pequeñas variaciones (variabilidad natural o causas comunes). Un proceso que funciona con sólo causas fortuitas de variabilidad se considera bajo control estadístico. Un sistema de medición se dice entonces que se encuentra bajo control estadístico cuando la variación en las mediciones se debe sólo a causas comunes y no a causas especiales. En este caso la variabilidad del sistema de medición es pequeña comparada con la variabilidad del proceso y/o con los límites de especificación o de tolerancia. Las fuentes de variabilidad que no forman parte de las causas fortuitas se denominan *causas atribuibles*. Un proceso que funciona en presencia de causas atribuibles se considera fuera de control (Montgomery, 1991). El control estadístico de procesos permite identificar y monitorear cambios entre ambos estados del proceso. Los procesos se evalúan periódicamente para asegurar los requerimientos de buen funcionamiento. La evaluación de las propiedades estadísticas del proceso son documentados en planillas de registro de observaciones donde se anotan todas las características del sistema de medición (variable medida, operaciones, equipos de

medición, personal, tamaño de muestra, estándares, límites, etc.), los datos relevados, los diagramas obtenidos y las conclusiones elaboradas a partir de éste.

Para el control de procesos, esta versión de InfoStat permite obtener diagramas de control tanto para atributos discretos como continuos. Otras técnicas estadísticas útiles para modelar las relaciones entre variables que determinan la salida de un proceso de producción y la calidad del producto obtenido, pueden ser fácilmente implementadas desde módulos básicos de InfoStat tales como regresión lineal y no lineal, análisis de varianza (para el estudio de causas que afectan el sistema), análisis multivariado y series de tiempo. En control de calidad, el proceso bajo estudio puede ser considerado como el proceso generatriz de una población de valores de una o más variables aleatorias sobre la que se desea inferir en base a la información contenida en una muestra extraída desde esa población. Por ello, las herramientas de la Estadística Descriptiva que InfoStat provee en los submenúes **Medidas resumen** y **Tablas de frecuencia** pueden ser usadas para caracterizar a la muestra de un proceso. En estos submenúes, a partir de una muestra de registros de una característica de interés, es posible construir automáticamente histogramas de frecuencias, tablas de frecuencias y calcular medias, varianzas, rangos, percentiles, etc. Cuando existen conocimientos suficientes como para proponer modelos de distribución de probabilidades para la descripción del proceso en estudio (se conocen los parámetros del proceso o la distribución de la variable aleatoria en estudio), el usuario de InfoStat puede a través del submenú Probabilidades y Cuantiles calcular probabilidades para responder, desde la Teoría Estadística, preguntas de interés en control de calidad tales como: ¿cuál será la probabilidad de que un producto tomado al azar desde la línea de producción contenga al menos un defecto si es que el proceso funciona como es esperado?. Cuando los parámetros del proceso no son conocidos, el usuario puede recurrir a la utilización de diversas técnicas estadísticas para inferir sobre esos parámetros. InfoStat permite estimar valores de parámetros mediante la técnica de intervalos de confianza y realizar pruebas de hipótesis sobre valores esperados y varianzas para el caso de una muestra y permite comparar parámetros de más de un proceso cuando se dispone de más de una muestra.

Los diagramas o cartas de control, ofrecidos dentro de la aplicación Control de Calidad, permiten la estimación de los parámetros o estándares que gobiernan un proceso bajo control a partir de muestras preliminares, el control de productos en línea desde el monitoreo del proceso usando estándares conocidos, y la estimación de la capacidad del proceso. Un diagrama de control o carta de control es una gráfica donde los valores de la característica de calidad estudiada se disponen sobre el eje de las ordenadas para distintas muestras o para una misma muestra en distintos momentos de tiempo que se identifican sobre el eje de las abscisas. Tres líneas (parámetros del diagrama) acompañan la serie graficada: la *línea media* (trazada a nivel de la media de los valores de la serie para un estado bajo control) y las líneas correspondientes a los *límites inferior y superior de control* (límites entre los que se espera queden comprendidas casi la totalidad de las observaciones de un proceso bajo control). Puntos fuera de la región determinada por ambos límites sugieren que el proceso no está bajo control. Aún si los valores de la serie observada se encuentran entre estos límites, el proceso puede cuestionarse por no poseer un patrón de distribución aleatoria de los puntos en torno a la línea media. Valores sistemáticamente mayores o menores al esperado sugieren un proceso fuera de control.

Los límites son obtenidos a partir de la expresión Límite = media± *k**error estándar, donde *k* es un entero. Cuando *k*=3 los intervalos son conocidos como intervalos tres sigmas ya que usualmente se representa con la letra griega sigma a la desviación estándar de las medias de la característica o error estándar. La amplitud entre los límites de control depende de la variabilidad de la característica de calidad, del tamaño de muestra utilizado para construir los límites y de la confiabilidad requerida (3 sigmas, 6 sigmas, etc.). Diagramas así construidos se llaman usualmente Diagramas Shewart.

En otras oportunidades la línea central y los límites de control son externamente determinados. Por ejemplo escogidos por el administrador del proceso de acuerdo a estándares preestablecidos o calculados a partir de muestras preliminares registradas no para controlar el proceso sino para establecer dichos límites.

Cuando se utilizan las cartas de control, se deben considerar las observaciones que caen fuera de los límites tanto como la tendencia del conjunto de observaciones y los puntos que caen sobre la línea media. La inestabilidad del proceso puede ser identificada por cualquiera de estas condiciones.

Dos tipos de errores se pueden cometer al tomar una decisión sobre el estado del proceso a partir de un diagrama de control: 1) Error tipo I, cuando se rechaza que el proceso está bajo control pero en realidad se encuentra en dicho estado, 2) Error tipo II, cuando se acepta que el proceso está bajo control pero en realidad está fuera de control. La probabilidad de cometer error tipo I es fijada por el usuario al construir el diagrama si usa límites probabilísticos. InfoStat no calcula límites probabilísticos sino límites del tipo k–sigmas. Con límites del tipo k-sigmas esta probabilidad puede ser aproximada si se conoce la distribución subyacente de estadístico utilizado para construir los límites de control. Para una distribución normal de las componentes aleatorias, un intervalo 3-sigmas para la calidad media del proceso, producirá un riesgo de error tipo I de 0.0027. La probabilidad de cometer error tipo II sólo puede ser calculada si se conoce la distribución del proceso cuando no está bajo control. Las curvas características de operación permiten visualizar la probabilidad de error tipo II bajo distintos estados del proceso e indican la potencia del diagrama para identificar cambios de estados de diferentes magnitudes (Montgomery, 1991). El tamaño del error tipo II puede leerse desde curvas características de operación. Estas se encuentran disponibles en la versión *full* de la aplicación.

Menú APLICACIONES ⇒ CONTROL DE CALIDAD permite seleccionar: 1) **Diagramas de control para atributos**: **Proporción de defectos (p)**, **Cantidad de defectos (np)**, **Número de casos por unidad (c)**; 2) **Diagramas de control de variables**: **Para la media y rango (X-barra, R)**, **Para la media y desvío estándar (X-barra, S)**; 3) **Diagrama de Pareto** y 4) **Capacidad de proceso (cp y cpk)**.

En los diagramas o cartas de control, para variables o para atributos, el usuario puede excluir subgrupos de datos desde el análisis tantas veces como sea necesario hasta conseguir el diagrama correspondiente al proceso bajo control. Cuando se establecen diagramas de control a partir de datos preliminares ya que no se dispone de estándares, muchas veces es de interés utilizar los parámetros del diagrama en el control del proceso en línea o control de observaciones futuras. Para todos los tipos de diagramas producidos, InfoStat permite activar la casilla **Parámetros de diagrama de control conocidos** e ingresar el valor de los

parámetros del diagrama (por ejemplo aquellos estimados a partir de una muestra preliminar) para obtener una gráfica a partir de los valores en el archivo pero con la línea central y los límites de control ingresados al activar la opción correspondiente.

El análisis de capacidad de proceso permite investigar si un proceso sigue funcionando bajo las especificaciones. Este análisis se basa en la distancia entre los resultados observados y los valores nominales o esperados bajo distribución normal. El usuario de InfoStat podrá estimar capacidad de procesos cuando solicite diagramas de control para variables.

Diagrama de control para atributos

Menú APLICACIONES ⇒ CONTROL DE CALIDAD ⇒ DIAGRAMAS DE CONTROL PARA ATRIBUTOS, permite obtener diagramas de control para características de calidad que no son medidas en una escala cuantitativa. Estos diagramas son útiles para situaciones donde cada producto inspeccionado es clasificado como *conforme* (no defectuoso) o *disconforme* (defectuoso) con las especificaciones de calidad. Cuando las características de calidad son discretas, como en este caso, se denominan *atributos*. Dentro de esta opción, InfoStat permite obtener diagramas para la proporción o porcentaje de productos disconformes resultantes de un proceso de producción (**Proporción de defectos (p)**), para el número de disconformes producidos (**Cantidad de defectos (np)**) y el diagrama de control de disconformidades por unidad (**Número de casos por unidad (c)**).

La proporción de disconformes, denotada por *p*, es el cociente entre el número de productos defectuosos y el total de elementos producidos. Un artículo *disconforme* es aquel que no cumple con al menos una de las características de calidad que se evalúan simultáneamente. Cuando se analiza la proporción de productos no defectuosos en lugar de la proporción de los disconformes se obtiene un diagrama de control para el rendimiento del proceso. El *diagrama de control para la proporción de defectos o diagrama p* es un diagrama de control de *Shewart*.

El diagrama se sustenta en la distribución de la proporción de defectuosos (distribución Binomial). Para un proceso bajo control, se supone que *p* es la probabilidad de que una pieza sea defectuosa o no cumpla con las especificaciones. Se supone también que existe independencia entre las piezas producidas. Si la variable aleatoria *X* representa el número de defectuosos en una muestra de tamaño *n*, *X* se distribuye como una Binomial con parámetros *n* y *p*. El valor esperado (media) para el número de defectuosos es *np* y su varianza *np(1-p)*. El estimador muestral de *p* es $\hat{p} = \frac{X}{n}$, cuya media es *p* y su varianza *p(1-p)/n*. Las líneas de referencia para el diagrama de Shewart basado en la proporción *p* teórica (proporción de referencia) son los *límites superiores e inferiores de confianza* (LSC y LIC respectivamente) y se construyen de la siguiente manera:

$$LSC = p + k\sqrt{\frac{p(1-p)}{n}}$$

$$Linea\ central = p$$

$$LIC = p - k\sqrt{\frac{p(1-p)}{n}}$$

En el caso de desconocer el valor de la proporción de disconformes *p*, de un proceso bajo control, InfoStat graficará los estimadores muestrales de *p* obtenidos a partir de muestras subsecuentes de *n* unidades y construirá las líneas de referencia a partir de los valores observados (estimadores). Usualmente la fracción de disconformes del proceso bajo control es desconocida. Las líneas visualizadas en el diagrama de control producido automáticamente por InfoStat son calculadas a partir de las observaciones en el archivo y por tanto corresponden a límites de prueba. Estas se obtienen de la siguiente manera:

$$LSC = \hat{p} + k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$Linea\ central = \hat{p}$$

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

donde $\hat{p} = \frac{\sum_{i=1}^{m} \hat{p}_i}{m}$ es el promedio de las proporciones de defectuosos a través de *m* muestras de tamaño *n*, cada una con proporción de defectuosos denotada por $\hat{p}_i$. InfoStat requiere que el usuario ingrese el valor de *k* para la construcción de los límites, por defecto *k* es 3.

Si alguna muestra se encuentra fuera de los límites de control de prueba, usualmente, después de considerar las posibles causas de este evento, se descarta dicha muestra y se recalculan los límites de prueba. Este proceso se repite hasta que todos los puntos se encuentren entre los límites, y en tal momento se aceptan los límites de control de prueba para el uso actual (Montgomery, 1991). Las facilidades de InfoStat para activar y desactivar casos permiten ajustar los límites de prueba en forma sencilla. Posteriormente, activando la casilla **Parámetros de diagrama de control conocidos** e ingresando el valor de los parámetros del diagrama obtenidos desde muestras preliminares, es posible obtener una gráfica a partir de nuevos valores en el archivo pero con la línea central y los límites de control ahora conocidos.

Ejemplo 49: En una línea de producción de elásticos para autos se toman 30 muestras de tamaño 200 cada una y se registra el número de defectuosos por muestra. Los datos se encuentran en el archivo Diagrama_p, cuyas columnas se llaman "tamaño" y "disconformes".

Para obtener el diagrama, activar el menú APLICACIONES ⇒ CONTROL DE CALIDAD y seleccionar **Proporción de defectos (p)**. Aparecerá una venta llamada **Diagramas de**

Control donde se listan las variables del archivo. El usuario deberá seleccionar la variable que contiene el número de defectuosos (en este ejemplo, la columna "Disconformes") e ingresarla en la subventana **Nro. Defectuosos**, la columna del archivo que contiene el tamaño de cada muestra (en este ejemplo, "Tamaño") deberá pasarse a la subventana **Total sub-grupos**. De ser necesario se puede indicar una variable que contiene fechas o que indexa el tiempo de extracción de la muestra de alguna forma, para ser usada en el eje de las abscisas del diagrama en la opción **Tiempo (Opcional)**. Por defecto el diagrama asume que el orden cronológico del muestreo es el utilizado para ingresar los datos. Al **Aceptar**, se visualizará otra pantalla denominada **Diagramas de Control** en la que el usuario puede modificar el valor de *k*. En la siguiente figura, se muestra el diagrama obtenido para *k*=3. En la ventana **Resultados** se listarán los valores de los límites superior e inferior y de la línea central.

Tabla 68: Límite de Control para la proporción de defectos. Archivo Diagrama_p.

Proporción defectos Límites de Control	
Línea Superior:	0.1382
Línea Central:	0.0805
Línea Inferior:	0.0228

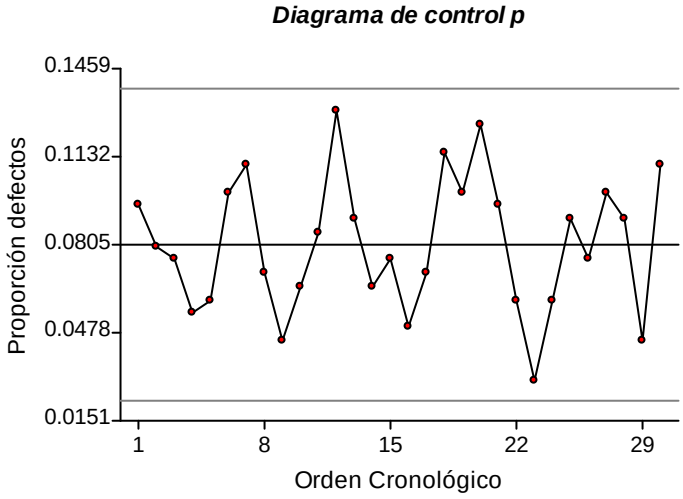


Figura 78: Diagrama de control p. Archivo Diagrama_p.

Si se tiene información sobre *p*, digamos que se especifica un valor objetivo de *p*=0.05, el LSC será $LSC = 0.05 + 3\sqrt{\frac{0.05(1-0.05)}{200}} = 0.0962$ y el LIC=0.0038. En la ventana **Diagramas de Control**, activando la casilla **Parámetros de diagrama de control conocidos** e ingresando el valor de los mismos se podrá obtener una gráfica a partir de los valores en el archivo pero con la línea central y los límites de control obtenidos a partir de la proporción *p* conocida.

En algunas oportunidades la fracción de disconformes de cada muestra es calculada sobre tamaños muestrales diferentes. InfoStat permite ingresar tamaños muestrales distintos para cada muestra y en tal caso basa los límites de control en un tamaño muestral promedio, calculado automáticamente. Este accionar produce límites aproximados. Duncan (1974) discute el cálculo del tamaño muestral necesario para tener una probabilidad alta de detectar un cambio en alguna cantidad especificada en el proceso.

El **diagrama de control (nP)** o **diagrama de cantidad de defectos (nP)**, se construye de manera análoga al diagrama anterior, pero en vez de graficar la fracción de disconformes se grafica directamente el número de disconformes. Las líneas de referencia del diagrama obtenidas a partir de los datos son calculadas de la siguiente manera:

$$LSC = n\hat{p} + k\sqrt{n\hat{p}(1-\hat{p})}$$
$$Linea\ central = n\hat{p}$$
$$LIC = n\hat{p} - k\sqrt{n\hat{p}(1-\hat{p})}$$

En la siguiente figura se muestra el diagrama de **Cantidad de defectos (nP)** para el ejemplo anterior (archivo *Diagrama_p*).

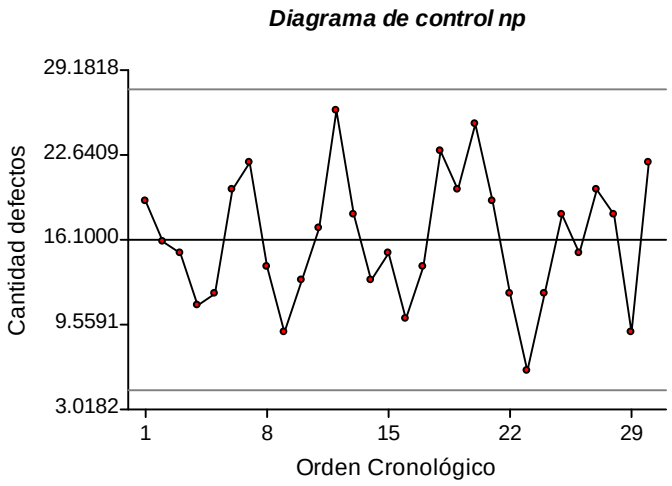


Figura 79: Diagrama de control np. Archivo Diagrama_p.

Otro diagrama que puede ser seleccionado es el diagrama de control **Número de casos por unidad (c)**. Este es útil para situaciones donde la muestra incluye varias unidades de inspección y sobre cada una de esas unidades se cuenta el número de disconformidades.

Ejemplo 50: Si un productor de videojuegos observa 50 muestras de n=10 videos cada una y cuenta en cada video el número de defectos, recolectará un conjunto de 50 promedios de no conformidades por unidad de inspección (video). Archivo Diagrama_c.

Si se ingresa en InfoStat la columna "Fallas/Video" y solicita diagrama **Número de casos por unidad (c)**, obtendrá una gráfica donde los límites serán calculados a partir de la distribución Poisson de la siguiente manera:

$$LSC = \hat{u} + k\sqrt{\hat{u}/n}$$
$$Linea\ central = \hat{u}$$
$$LIC = \hat{u} - k\sqrt{\hat{u}/n}$$

donde $\hat{u}$ representa el promedio a través de todas las muestras del número promedio de no conformidades por unidad. Si el archivo de datos contiene el número de disconformidades para cada unidad de inspección, primero se deberá calcular, usando el menú **Medidas resumen**, el número promedio de defectos por unidad para cada muestra. El archivo de promedios resultante será utilizado para producir el diagrama de control c.

Para obtenerlo, activar el menú APLICACIONES ⇒ CONTROL DE CALIDAD ⇒ DIAGRAMAS DE CONTROL PARA ATRIBUTOS y seleccionar **Número de casos por unidad (c).** Aparecerá una ventana llamada **Diagramas de Control** donde se listan las variables del archivo. El usuario deberá seleccionar la variable que contiene el número promedio de defectuosos por unidad de inspección (en este ejemplo, la columna "Fallas/Video") e ingresarla a la subventana **Nro. de Casos por Unidad**. Opcionalmente se puede indicar una variable que contiene fechas o que indexa el tiempo de extracción de la muestra de alguna forma, para ser usada en el eje de las abscisas del diagrama en la subventana **Tiempo (Opcional)**. Por defecto, el diagrama asume que los datos del muestreo fueron ingresados en orden cronológico. Al **Aceptar**, se visualizará otra pantalla denominada **Diagramas de Control** en la que el usuario puede modificar el valor de *k*. En la siguiente figura, se muestra el diagrama obtenido para *k*=3. En la ventana **Resultados** se listarán los valores de los límites de prueba y de la línea central.

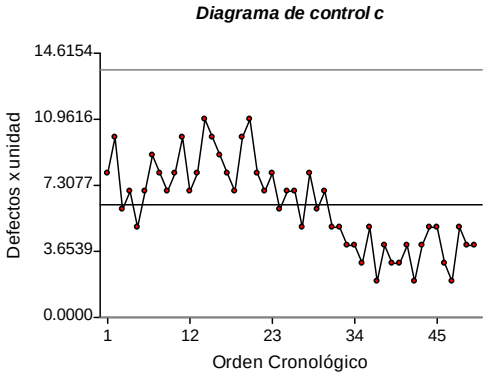


Figura 80: Diagrama de control c. Archivo Diagrama_c.

El gráfico muestra que el proceso no se encuentra bajo control.

Diagramas de control de variables

Menú APLICACIONES ⇒ CONTROL DE CALIDAD ⇒ Diagramas de control de variables, permite obtener diagramas de control para características de calidad que son

medidas en una escala cuantitativa o numérica. Estos diagramas son útiles para situaciones donde de cada producto inspeccionado se obtiene una medida para una característica de calidad. En tales situaciones interesa controlar el valor medio de la característica y alguna medida de su variabilidad. Dentro de esta opción, InfoStat permite obtener los siguientes diagramas: **Para la media y rango** (X-barra, R) y **Para la media y desvío estándar (X-barra, S)** . InfoStat construye los diagramas bajo el supuesto de distribución normal para la característica de calidad en estudio. Los diagramas resultantes son aproximadamente correctos para distribuciones no normales cuando se trabaja con muestras grandes.

Los diagramas para la media y rango permiten el control de la calidad media del proceso (**diagrama X-barra**) y de la variabilidad a través de la amplitud o rango del mismo (**diagrama R**). Estos proporcionan información sobre la capacidad de funcionamiento del proceso, mientras que el diagrama X-barra sirve para controlar la variabilidad entre muestras o variabilidad en el proceso con el tiempo, el diagrama R permite monitorear la variabilidad dentro de una muestra o variabilidad instantánea del proceso. Generalmente se usan de 20 a 30 muestras preliminares, tomadas cuando el sistema se supone bajo control, para obtener los valores de μ y σ, la media y desviación estándar del proceso respectivamente. Por ejemplo, si se dispone de m=20 muestras de tamaño n=5 cada una, el estimador de μ, el cual representa la línea central del diagrama, será:

$$\hat{u} = \frac{\overline{X}_1 + ... + \overline{X}_{20}}{20}$$

y el estimador del rango es calculado a partir de los rangos muestrales como:

$$\hat{R} = \frac{R_1 + ... + R_{20}}{20}$$

donde R_i es el rango, o diferencia entre el máximo y mínimo valor, en la muestra i-ésima.

InfoStat construye el diagrama de control de media (diagrama X-barra) usando las siguientes líneas de referencia:

$$LSC = \hat{u} + k\hat{\sigma}_{\hat{\mu}}$$
$$Linea\,central = \hat{u}$$
$$LIC = \hat{u} - k\hat{\sigma}_{\hat{\mu}}$$

donde LSC y LIC son los límites superior e inferior de confianza respectivamente y $\hat{\sigma}_{\hat{\mu}}$ es calculado a partir del rango como $\hat{\sigma}_{\hat{\mu}} = \frac{1}{\sqrt{n}}\frac{\hat{R}}{d_2}$ siendo los valores de d_2 aquellos correspondientes a la esperanza de la variable aleatoria $\frac{R}{\sigma}$ o amplitud relativa (Montgomery, 1991).

InfoStat representa también los valores de R de muestras sucesivas en el diagrama de control de rango (diagrama R) usando las siguientes líneas de referencia:

$$LSC = \hat{R} + k\hat{\sigma}_R$$
$$Linea\ central = \hat{R}$$
$$LIC = \hat{R} - k\hat{\sigma}_R$$

donde el estimador $\hat{\sigma}_R$ se obtiene también a partir de la distribución de la amplitud relativa, $\frac{R}{\sigma}$. La desviación estándar de $\frac{R}{\sigma}$, denotada por d_3, es una función conocida de n. Los valores de d_3 se encuentran tabulados en Montgomery (1991) para distintos tamaños de muestra. InfoStat calcula a partir de ellos $\hat{\sigma}_R$ como $\hat{\sigma}_R = d_3 \frac{\hat{R}}{d_2}$.

Ejemplo 51: El archivo Diagrama_MyR contiene lecturas de diámetros de anillos de pistón forjados para 25 muestras de 5 anillos cada una (Montgomery, 1991).

Para obtener el diagrama, activar el menú APLICACIONES ⇒ CONTROL DE CALIDAD ⇒ DIAGRAMAS DE CONTROL DE VARIABLES y seleccionar: **Para la media y el rango (X-barra, R)**. Aparecerá una venta llamada **Diagramas de Control** donde se listan las cinco variables del archivo, cada una representando una observación muestral. El usuario deberá seleccionar todas las columnas que contienen observaciones muestrales ("Obs1-Obs5") e ingresarlas en la subventana **Observaciones por muestra**. Opcionalmente se puede indicar una variable que contiene fechas o que indexa el tiempo de extracción de la muestra de alguna forma, para ser usada en el eje de las abscisas del diagrama. Por defecto el diagrama asume que el orden cronológico del muestreo es el utilizado para ingresar los datos. Al **Aceptar**, se visualizará otra pantalla denominada **Diagramas de Control** en la que el usuario puede modificar el valor de *k*. En las siguientes figuras, se muestran los diagramas (para la media y para el rango) obtenidos para *k*=3. En la ventana **Resultados** se listarán los valores de media y rango para cada muestra y los valores de la línea central y los límites de control de los diagramas para la media y para el rango.

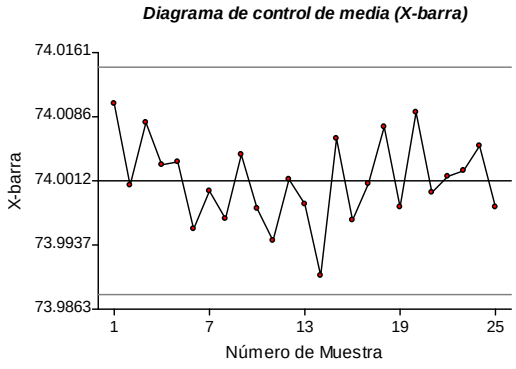


Figura 81: Diagrama de control de media (x-barra). Archivo Diagrama_MyR.

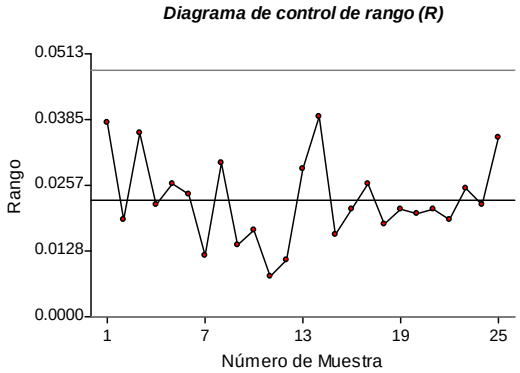


Figura 82: Diagrama de control de rango (R). Archivo Diagrama_MyR.

La media de las medias muestrales es 74.001 (línea central del diagrama X-barra) y la media de los rangos muestrales es 0.023 (línea central del diagrama R). Los valores de d_2 y d_3 para n=5 son 2.326 y 0.864, respectivamente (Montgomery, 1991). Los límites para los diagramas de medias y rangos para k=3 son:

$$74.001 \pm 3\frac{0.023}{2.326\sqrt{5}} \quad \text{y} \quad 0.023 \pm 3 * 0.864\frac{0.023}{2.326}.$$

Dado que en el diagrama R se visualiza que la variabilidad del proceso está bajo control se observa el diagrama X-barra. A partir de este último se informa que no hay evidencias de falta de control para el nivel de calidad promedio del proceso. Estos límites de control podrían ser adoptados para el control en línea del proceso. Si en el diagrama R se visualizan puntos fuera de control, primero deberán recalcularse los límites de control eliminando las causas atribuibles y recién después realizar interpretaciones a partir del diagrama X-barra.

InfoStat también permite obtener **diagramas de media y desviación estándar**. Los diagramas X-barra y S se utilizan cuando los tamaños de muestra son relativamente grandes (n=10 o más observaciones por muestra) ya que en estos casos el rango (R) no realiza un uso eficiente de la información. La raíz cuadrada del estimador insesgado de $\hat{\sigma}^2$ es utilizado para obtener S en cada muestra de n observaciones:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Dado que el valor esperado de S es $c\sigma$ y la desviación estándar de S es $\sigma\sqrt{1-c^2}$ donde c es una constante que depende del tamaño muestral, los límites de control para el diagrama S cuando se dispone de un valor estándar para σ son los siguientes:

$$LSC = c\sigma + k\sigma\sqrt{1-c^2}$$
$$Linea\ central = \sigma$$
$$LIC = c\sigma - k\sigma\sqrt{1-c^2}$$

donde $c = \left(\frac{2}{n-1}\right)^{1/2} \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}$. Si se desconoce σ, los límites son calculados a partir del conjunto de *m* muestras de prueba usando el estadístico $\hat{S}/c$, el cual es un estimador insesgado de σ con $\hat{S} = \frac{1}{m}\sum_{i=1}^{m} S_i$ y S_i la desviación estándar muestral de la muestra i-ésima. Luego, los límites de control son calculados de la siguiente manera:

$$LSC = c\hat{S} + k\hat{S}\sqrt{1-c^2}$$
$$Linea\ central = \hat{S}$$
$$LIC = c\hat{S} - k\hat{S}\sqrt{1-c^2}$$

Los límites de control para el diagrama X-barra, son definidos a partir del estimador de la desviación estándar muestral como:

$$LSC = \hat{\mu} + k\frac{\hat{S}}{c\sqrt{n}}$$
$$Linea\ central = \hat{\mu}$$
$$LIC = \hat{\mu} - k\frac{\hat{S}}{c\sqrt{n}}$$

El control en línea de un proceso puede realizarse activando la casilla **Parámetros de diagrama de control conocidos** (subventana **Diagramas de Control**) e ingresando el valor de los parámetros del diagrama obtenidos desde muestras preliminares, a partir de la media del proceso y su desviación estándar (calculada como $\hat{\sigma} = \frac{\hat{R}}{d_2}$).

Para el archivo *Diagrama_MyR* se obtuvieron los diagramas X-barra y S invocando el menú APLICACIONES ⇒ CONTROL DE CALIDAD ⇒ DIAGRAMAS DE CONTROL DE VARIABLES y seleccionando **Para la media y desvío estándard (X-barra, S)**. Aparece una venta llamada **Diagramas de Control** donde se listan las variables del archivo, cada una representando una observación muestral. Se seleccionaron todas las columnas que contienen observaciones muestrales ("Obs1-Obs5") y se incluyeron en la subventana **Observaciones por muestra**. Opcionalmente se puede indicar una variable que contiene fechas o que indexa el tiempo de extracción de la muestra de alguna forma, para ser usada en el eje de las abscisas del diagrama. Por defecto el diagrama asume que el orden cronológico del muestreo es el utilizado para ingresar los datos. Al **Aceptar**, se visualiza

otra pantalla denominada **Diagramas de Control** en la que el usuario puede modificar el valor de *k*. En la siguiente figura, se muestra el diagrama de control de desvío estándar obtenido para *k*=3 (el diagrama X-barra no se muestra aquí, pero se obtiene automáticamente y se interpreta de la misma manera que el del ejemplo anterior). En la ventana **Resultados** se listarán los valores de media y S para cada muestra y los valores de la línea central, los límites de control y el análisis de capacidad del proceso. Los resultados permiten arribar a las mismas conclusiones que el ejemplo anterior y son muy similares debido a que el ejemplo involucra muestra de tamaño *n*=5 donde el uso del rango es apropiado.

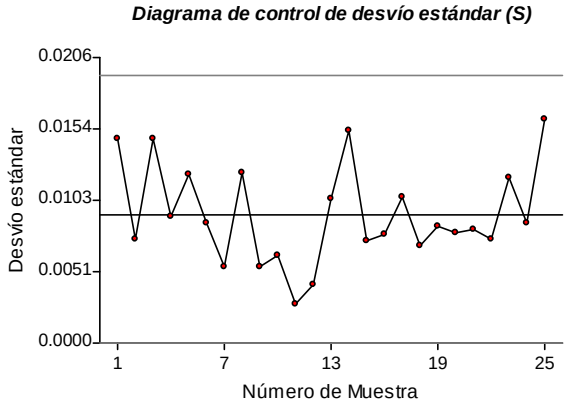


Figura 83: Diagrama de control de desvío estándar (S). Archivo Diagrama_MyR

Diagrama de Pareto

Menú APLICACIONES ⇒ CONTROL DE CALIDAD ⇒ DIAGRAMA PARETO permite obtener un diagrama de barra donde se muestran las frecuencias relativas de los distintos tipos de errores que se pueden detectar en un proceso de control de calidad. La característica particular de este diagrama es que los errores se muestran ordenados conforme su frecuencia de mayor a menor. En muchos procesos se controla el número de errores de distintos tipos sobre cada unidad de inspección ya que la importancia de los mismos puede ser diferente. Para realizar este diagrama se deberá tener un archivo donde los distintos tipos de errores son tratados como variables (columnas del archivo) y cada registro corresponde a una unidad de inspección. Los valores a ingresar en cada celda pueden ser 0,1,2,3,... y representan el número de errores del tipo en cuestión para cada unidad.

Ejemplo 52: El archivo Apareto contiene registros de 6 tipos de errores realizados por dos operarios sobre un total de 64 piezas inspeccionadas.

Para obtener el diagrama, activar el menú APLICACIONES ⇒ CONTROL DE CALIDAD ⇒ DIAGRAMA DE PARETO y seleccionar las 6 columnas conteniendo los tipos de error como variables. Si se selecciona la variable operario como criterio de clasificación (opcional) InfoStat reportará que proporción del total de cada tipo de error fue contabilizado por cada operario. En el gráfico que se muestra a continuación se seleccionaron los 6 tipos de error y no se clasificó por operario.

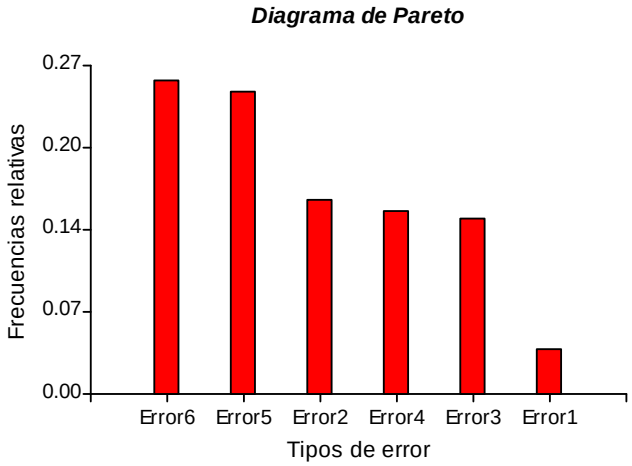


Figura 84: Diagrama de Pareto. Archivo Apareto

Capacidad de Proceso

InfoStat informa sobre la **Capacidad del proceso (cp y cpk)** calculando las proporciones esperadas de productos fuera de especificación. El usuario debe proveer límites de especificación en la subventana **Diagramas de control** en el campo **Límites de especificación (Capacidad de proceso)** que se presenta luego de seleccionar las columnas del archivo que intervienen en el análisis. La proporción de productos fuera de especificación es la suma de las proporciones esperadas de productos con mediciones inferiores al límite inferior de especificación y mayores al límite superior. Estas proporciones son estimadas a partir del modelo normal usando la media y desviación estándar estimada del proceso bajo control:

$$\text{Inferior}: P(X < LIE) = \Phi^{-1}\left(Z = \frac{LIE - \hat{\mu}}{\hat{\sigma}}\right)$$

$$\text{Superior}: P(X > LSE) = 1 - \Phi^{-1}\left(Z = \frac{LSE - \hat{\mu}}{\hat{\sigma}}\right)$$

$$\text{Total}: P(X < LIE) + P(X > LSE)$$

donde LIE y LSE son los límites inferior y superior de especificación ingresados por el usuario (éstos son determinados externamente y son independientes de los límites de control). Los valores de la variable *Z* usados para el cálculo de dichas probabilidades son conocidos como Tolerancias Bilaterales.

La Tabla 69 muestra las proporciones esperadas para el ejemplo del archivo *Diagrama MyR*, tal como son reportadas en la ventana **Resultados** por InfoStat. La proporción total de productos fuera de los límites especificados (73.97 y 74.03 ya que se requieren anillos con diámetro 74.000±0.03 mm) es 0.0020, por lo que se concluye que el 0.20% de los anillos forjados tendrán diámetros distintos a los especificados. InfoStat muestra bajo el nombre de

tolerancias bilaterales a los valores de la variable normal estándar correspondiente a la estandarización de los límites inferior y superior de especificación. Por último, InfoStat proporciona la evaluación de la aptitud del proceso a través de las estadísticas *cp* y *cpk*. Estos índices suponen que los límites naturales de tolerancia del proceso son cercanos a los límites de especificación. Considerando que la amplitud entre los límites naturales es igual a 6 sigmas, se calcula el cociente entre la amplitud *LSE-LIE* y 6 sigmas. Se espera que la relación de capacidad de o aptitud de proceso, *cp*, sea ligeramente mayor que uno si el proceso está bajo control. Además InfoStat reporta los valores *cpk* correspondientes al valor absoluto del mínimo de las tolerancias bilaterales dividido 3 (Montgomery, 1991).

Tabla 69: Análisis de capacidad de proceso. Archivo Diagrama_MyR.

Diagramas de Control
Análisis de la capacidad del proceso

Item	Valor
Media:	74.00
Desv.est.:	0.01
LEI:	73.97
LES:	74.03

Tolerancias bilaterales

Item	Valor
Zinf:	-3.22
Zsup:	3.01

Proporciones fuera de especificación

Item	Valor
Inferior:	0.0006
Superior:	0.0013
Total:	0.0020

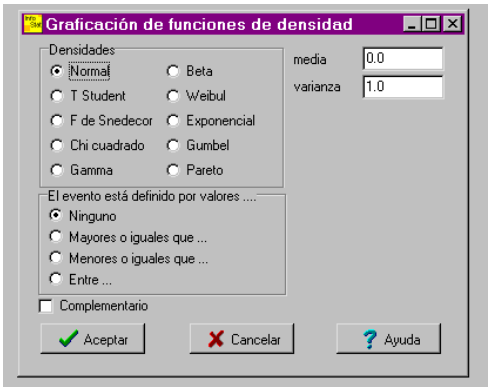
Evaluación aptitud del proceso

Item	Valor
cp	1.04
cpk	1.00

Aplicaciones Didácticas

Gráficos de funciones de densidad continuas

Menú APLICACIONES ⇒ DIDACTICAS ⇒ GRAFICOS DE FUNCIONES DE DENSIDAD CONTINUAS, permite obtener gráficos de funciones de densidad y visualizar las distintas formas que adoptan esas densidades al modificar los parámetros que las caracterizan. También es posible sombrear y obtener el tamaño de áreas bajo la curva definida por la función de densidad. Así, es posible obtener en forma automática probabilidades asociadas a distintos eventos de interés. Al acceder al menú aparecerá la siguiente ventana:

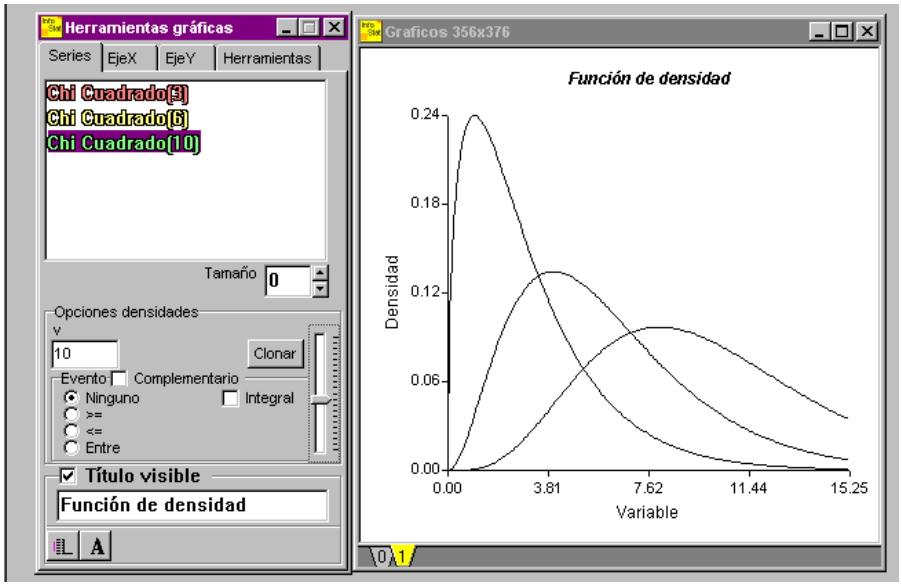


En esta ventana se deberá seleccionar la densidad de interés, habilitándose los campos para ingresar los parámetros que las caracterizan. Para consultar características acerca de las distribuciones disponibles véase en el Capítulo Manejo de Datos correspondiente al menú DATOS, submenú LLENAR CON..., del presente manual. La subventana **El evento está definido por valores...** se debe utilizar para definir el evento aleatorio para el cual se desea asignar una medida de probabilidad. La definición del evento también se puede realizar luego de haber generado la gráfica de la distribución desde la ventana de **Herramientas gráficas**.

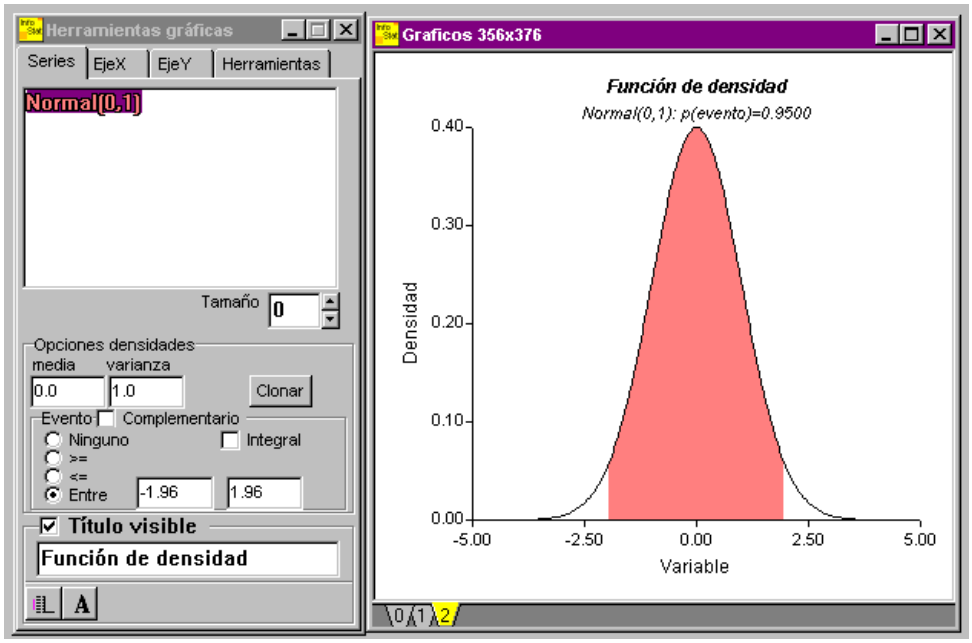
Para dar un ejemplo del uso del graficador supongamos que se elige la densidad Chi cuadrado. Al activar **Chi cuadrado** aparecerá el campo para que se ingresen los grados de libertad (el único parámetro que caracteriza la distribución). Después de indicar los grados de libertad de la distribución y presionar **Aceptar**, aparecerá el gráfico de la distribución acompañado por la ventana **Herramientas gráficas**. Dicha ventana permite modificar los atributos del gráfico. En la Solapa **Series** se mostrarán los campos correspondientes a los parámetros de la distribución usada, en este caso se muestra el campo **v** (los grados de libertad con los que se generó la distribución). Si el valor de este campo es modificado, automáticamente se modificará el gráfico mostrando la distribución para el nuevo valor del parámetro ingresado desde la ventana **Herramientas Gráficas**. La modificación del parámetro puede hacerse ingresando un nuevo valor en el campo correspondiente o utilizando la regla de desplazamiento que aparece automáticamente al hacer un *click* sobre el campo que contiene el valor del parámetro.

Para observar distintas distribuciones (series gráficas) sobre un mismo gráfico se dispone del botón **Clonar**. Cuando existe una gráfica de distribución en la ventana **Gráficos** y se usa el botón **Clonar** se obtendrá automáticamente una copia (clon) de la serie gráfica original en la solapa **Series**. Activando la nueva serie (desde la solapa **Series**) y cambiando el valor de sus parámetros es posible visualizar ambas distribuciones a la vez. Este procedimiento puede repetirse tantas veces como series gráficas (funciones de densidad) el usuario desee visualizar simultáneamente sobre la misma ventana gráfica.

A continuación se presenta una ventana gráfica mostrando tres distribuciones Chi cuadrado con 3, 6 y 10 grados de libertad respectivamente. Para obtener este gráfico se solicitó una Chi cuadrado con 3 grados de libertad **(v)**, luego se clonó presionando el botón **Clonar** y seleccionando la serie clonada se indicó el valor 6 en el campo **v**; finalmente se clonó esta última y se cambió el número 6 por 10 en el campo **v**.

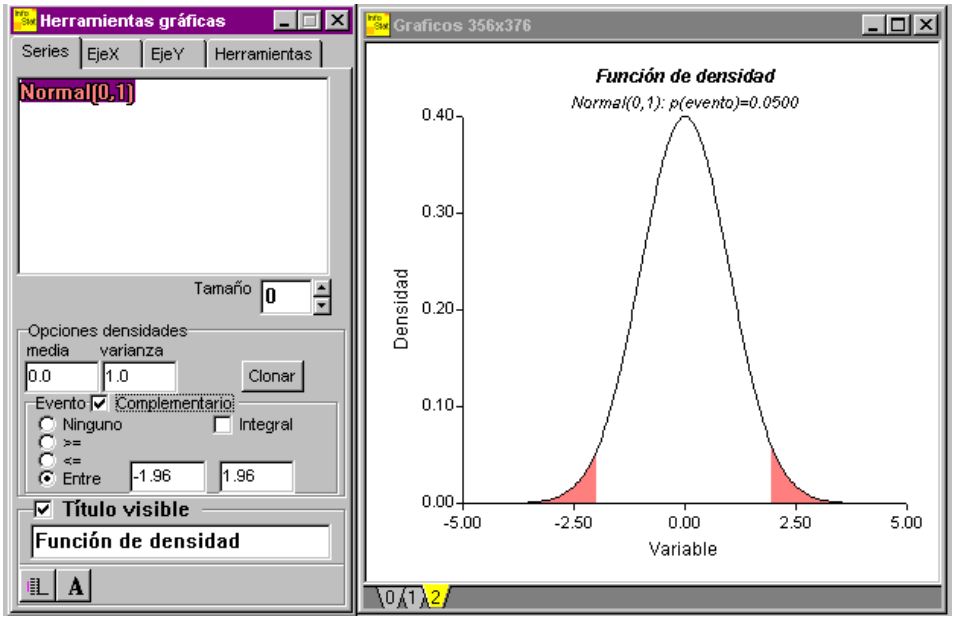


La subventana **Evento** permite definir un evento aleatorio para visualizar el área correspondiente a la probabilidad de ocurrencia del mismo. Se deberá activar la casilla menor o igual (< =), mayor o igual (> =) o entre y se deberá ingresar el o los valores que definen el evento. Por ejemplo, si el evento corresponde a valores de una variable aleatoria normal estándar entre –1.96 y 1.96, habiendo seleccionado la densidad normal con media=0 y varianza=1, se activará en la subventana **Evento** la opción **Entre** ingresando en los dos campos que aparecerán los valores –1.96 y +1.96 Se obtendrá un gráfico como el que se presenta a continuación:



En el gráfico generado se puede leer la probabilidad del evento de interés, en este caso *p(evento)=0.9500.* Si se activa la casilla **Complementario** se podrá leer la probabilidad del evento complementario es decir aquel que comprende valores de la variable menores que –1.96 y valores mayores que +1.96. La región sombreada corresponderá al evento complementario.

El campo **Integral**, en la ventana Herramientas Gráficas para funciones de densidad continuas, permite visualizar la función de distribución (o distribución acumulada) de la densidad correspondiente a la serie seleccionada. Al activar este campo la densidad representada gráficamente es reemplazada por el integral de la misma en el dominio de valores de la variable aleatoria en estudio.



Uso del graficador de funciones de densidad para la visualización de conceptos sobre prueba de hipótesis

Esta aplicación de InfoStat puede ser usada para ilustrar el procedimiento de prueba de hipótesis.

Supongamos que se tiene una densidad que depende de un parámetro desconocido y se toma una muestra aleatoria $(X_1,...,X_n)$ de tamaño *n*. El problema de contraste de hipótesis, *i.e.* un problema para el cuál existen sólo dos acciones posibles (aceptar o rechazar la hipótesis planteada), implica decidir (a partir del valor de un estadístico calculado desde la muestra disponible) si el parámetro desconocido de la densidad, digamos θ, es igual, menor o mayor que un valor especificado arbitrariamente, digamos θ_0 (se supone que la hipótesis es sobre un parámetro de la distribución).

Por ejemplo, si se ha diseñado un nuevo tratamiento para una enfermedad se podría contrastar la hipótesis que establece que la respuesta promedio de pacientes que reciben el

nuevo tratamiento es mayor a un valor θ_0 que representa la respuesta promedio de pacientes que no reciben el nuevo tratamiento (pacientes de un grupo control).

Un problema más simple es considerar la situación donde se quiere probar que el parámetro de la distribución es igual a un valor θ_0 o un valor θ_1 (se supone que el parámetro θ asume uno de estos dos valores). Se tienen entonces dos hipótesis:

$$H_0: \theta = \theta_0 \text{ y } H_1: \theta = \theta_1$$

Concluido el experimento a partir del valor observado en el estadístico de la prueba se rechazará o no H_0. El espacio muestral para la prueba de hipótesis deberá ser dividido en dos partes: R_0 y R_1. Se denota con R_0 a la región asociada a la aceptación de H_0 y con R_1 a la región remanente asociada al rechazo de H_0. Esto significa que si una muestra aleatoria produce un valor del estadístico de la prueba que pertenece a R_0 pensaremos que la hipótesis H_0 es la correcta. En caso contrario rechazaremos H_0 y nos quedaremos con la hipótesis H_1.

Debido a que se tomará una decisión a partir de una muestra aleatoria, se debe tener presente que existe la chance de cometer error. Un error (error tipo I) puede producirse cuando a pesar que H_0 sea verdadera (cuando presupone como valor del parámetro el valor del mismo en la población), la muestra produce un valor del estadístico de la prueba que pertenece a R_1 y por tanto decidiremos incorrectamente a favor de H_1. El otro error posible (error tipo II) se refiere a situaciones donde la hipótesis verdadera es H_1 pero la muestra produce un valor de estadístico de la prueba que pertenece a R_0, por lo que no se rechaza H_0 y por tanto decidiremos incorrectamente a favor de H_0.

Las probabilidades de realizar acciones incorrectas son denotadas por:

α= probabilidad de rechazar H_0 cuando ha sido postulada correctamente

β= probabilidad de no rechazar H_0 cuando la hipótesis correcta es H_1.

Las probabilidades de cometer error son conocidas como tamaño del error, α es el tamaño de error tipo I y β es el tamaño de error tipo II. Como las acciones a tomar son excluyentes, en un experimento particular, sólo puede cometerse error de tipo I o error de tipo II.

Tradicionalmente la región R_1 es conocida como región crítica o de rechazo para la prueba de H_0. La probabilidad de obtener un punto muestral que pertenece a la región crítica cuando H_0 es verdadera, es decir α, también suele ser denominada tamaño de la región crítica.

El problema relacionado a contrastar la hipótesis nula es encontrar una región crítica de tamaño α que minimice β. Vía el lema de Neyman-Pearson, se deriva un método para obtener la región crítica de tamaño α que minimiza la probabilidad β respecto a todas las regiones críticas cuyo tamaño no excede α (mejor región crítica). La mejor prueba de hipótesis es aquella basada en la mejor región crítica. Luego el problema se reduce a la identificación de los valores que delimitan o definen dicha región (valores o puntos críticos).

La probabilidad 1-β de rechazar una hipótesis nula falsa es conocida como potencia de la prueba. La potencia de la prueba será menor a medida que crece β. InfoStat permite ilustrar cómo disminuye β, conforme se aumenta el tamaño muestral, el valor α o la distancia entre los valores del parámetro especificados bajo H_0 y H_1 (hipótesis nula e hipótesis alternativa).

Como ilustración numérica considérese el problema de probar si la media de una variable aleatoria en la población es 50 o 52 a partir de los datos de una muestra aleatoria de tamaño 25. Supóngase que se conoce que la variable aleatoria se distribuye normal con varianza σ^2=100 y también supóngase que la media de la muestra obtenida, $\bar{X}$, es 54. Luego las hipótesis a contrastar son: H_0: θ=50 y H_1: θ=52 donde el parámetro θ representa la media de la variable en estudio, en este ejemplo $\theta=\mu$.

La región crítica R_1 queda definida por los valores de $\bar{X} \geq c$, donde *c* es elegido de manera tal que $P(\bar{X} \geq c | \mu=50)=\alpha$. Tomando α=0.05, el valor *c* puede ser obtenido en InfoStat de la siguiente manera:

Generar la distribución del estadístico bajo la hipótesis nula. Esto es una normal con parámetros media=50 y varianza=4, dado que si X se distribuye normal media=50 y varianza=100, el estadístico $\bar{X}$ se distribuirá normal con media μ=50 y varianza 100/25=4.

En **El evento está definido por valores...** activar la opción **Mayores o iguales que...**, aparecerá el punto crítico *c*, ya que InfoStat reporta automáticamente el cuantil 0.95 de la distribución al activar la opción. Luego *c*=53.28 es el punto que delimita las regiones R_1 y R_0. En la ventana **Gráficos** se visualizará la distribución y el área sombreada correspondiente a la probabilidad del evento.

Si se desean obtener regiones críticas de otro tamaño el usuario deberá ingresar el valor crítico correspondiente en el campo que se habilita al definir el evento.

El punto crítico es aquel que satisface la ecuación:

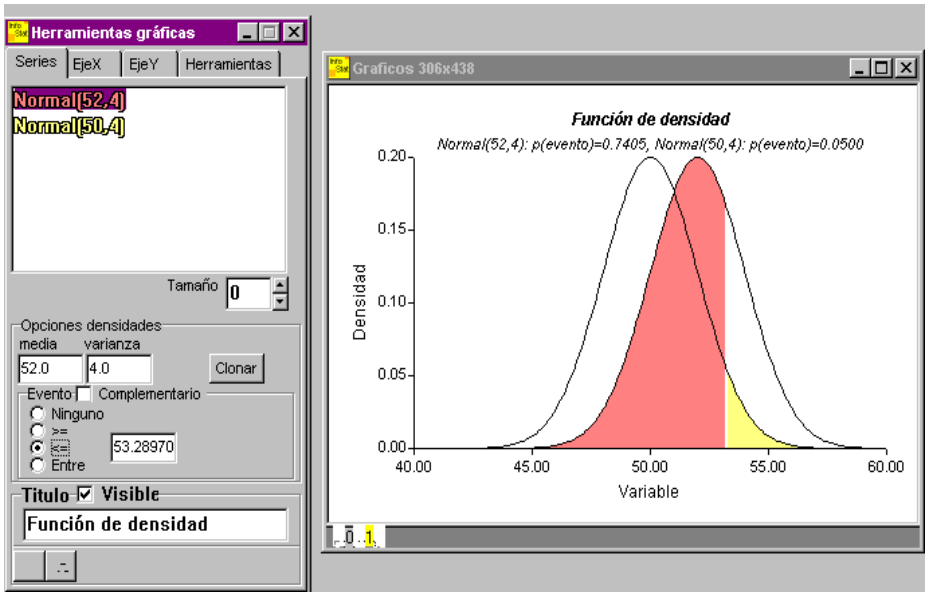
$$Z = \frac{(c-50)}{\sqrt{100/25}} = 1.645$$

La región crítica corresponde a los puntos muestrales para los cuales $\bar{X} \geq 53.28$. Luego si el valor observado de la media muestral es 54 ésta pertenece a la región de rechazo y se deberá rechazar H_0 a favor de H_1.

Consideramos ahora el problema numérico de calcular β, asumiendo μ_0=50 y μ_1=52, *n*=25, varianza σ^2=100 y α=0.05. Recordemos que $\beta=P(\bar{X} \in R_0 | H_1)$, la probabilidad asociada al evento "el estadístico pertenece a la región de aceptación dado que la hipótesis nula es falsa". Luego, $\beta=P(\bar{X} \leq 53.28 | \mu=52)$. Para obtener el valor de β en InfoStat se podrían seguir los siguientes pasos:

Sobre la gráfica anterior generar la distribución del estadístico $\bar{X}$ bajo la hipótesis alternativa. Es decir graficar una densidad normal con parámetros media=52 y varianza=4. Para lograr esto se deberá **Clonar** la serie gráfica existente y cambiar el parámetro media ingresando 52, tarea realizada desde la ventana **Herramientas gráficas**.

En **Evento** activar la opción <= y en el campo escribir 53.28. La porción sombreada de esta distribución corresponde a β. Se puede leer debajo del título del gráfico, el valor de la probabilidad de error de tipo II como *p(evento)=0.7405*.



Los ejemplos presentados ilustran como se puede utilizar InfoStat para trabajar el significado geométrico de α y β. El mismo ejemplo podría haber sido abordado desde la distribución del estadístico Z. Usando otras densidades se podría ejemplificar el procedimiento de prueba de hipótesis para estadísticos asociados a la varianza u otro parámetro desconocido de la distribución.

Intervalos de confianza

Menú APLICACIONES ⇒ DIDACTICAS ⇒ INTERVALOS DE CONFIANZA, permite obtener por simulación un conjunto de intervalos de confianza para la esperanza de una variable aleatoria normal. El propósito de este procedimiento es la visualización empírica del concepto de *confianza* subyacente en el procedimiento de estimación por intervalos de confianza.

El método de estimación por intervalos de confianza determina la exactitud en la estimación del parámetro de interés. El método generalmente es descrito a través del clásico ejemplo de estimación de la media μ de una densidad normal con varianza conocida σ^2, a partir de una muestra de tamaño *n*.

Para este caso $\bar{X}$ es el mejor estimador insesgado de μ, por lo que la estimación será basada en $\bar{X}$, con distribución normal con media μ y varianza σ^2/n. Esta distribución permite calcular la probabilidad de hallar valores de $\bar{X}$ en un intervalo específico en torno de un parámetro μ. También se conoce que la variable $Z = \dfrac{\left(\bar{X} - \mu\right)}{\sqrt{\sigma^2/n}}$ tiene distribución normal estándar, por lo que P(|Z| < 1.96) =0.95.

Luego $P\left(-1.96 \leq \frac{\left(\bar{X}-\mu\right)}{\sqrt{\sigma^2/n}} \leq 1.96\right) = 0.95$.

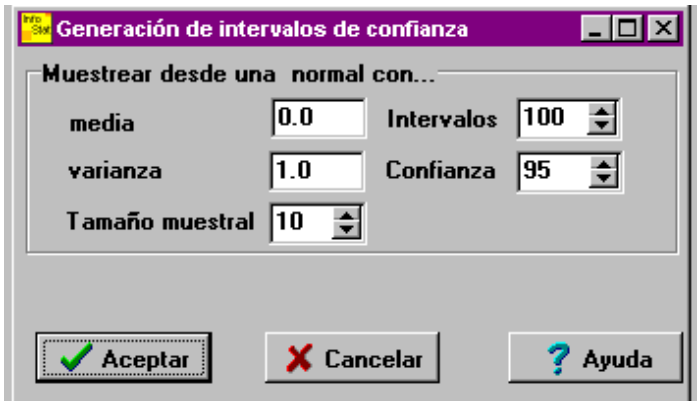
Reordenando se tiene que:

$$P\left(\bar{X} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{\sigma^2/n}\right) = 0.95$$

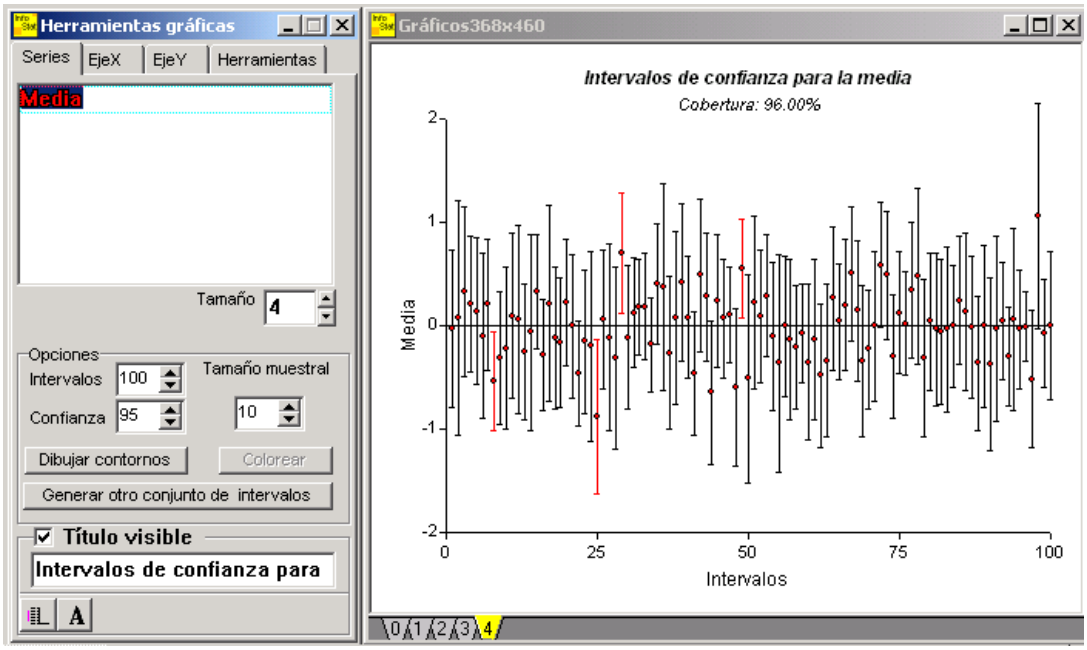
Si esta probabilidad es interpretada operacionalmente en términos de la frecuencia relativa del evento indicado a lo largo de muchas repeticiones del experimento de muestreo, el intervalo de confianza precedente establece que el 95% de los intervalos de la forma obtenida anteriormente, cada uno proveniente de muestras de tamaño *n*, contendrán la media μ en su interior.

El intervalo así logrado se conoce con el nombre de intervalo de confianza del 95% para μ y los valores mínimo y máximo del intervalo son conocidos con el nombre de límites de confianza para μ. InfoStat permite ilustrar el concepto operacional de probabilidad subyacente en la estimación del intervalo de confianza ya que, a partir de la distribución de la variable X definida por el usuario, obtiene por simulación un número *H* de muestras y para cada muestra calcula los límites del intervalo de confianza descriptos anteriormente. Los *H* intervalos obtenidos son dispuestos en una gráfica con fines comparativos. Sobre la misma gráfica InfoStat traza sobre el eje de las ordenadas el valor del parámetro (Media). Claramente en la práctica se dispone de sólo una muestra y no se conoce la media de la distribución, es justamente este parámetro el que se pretende estimar. Sin embargo en esta aplicación la distribución subyacente se supone conocida (la media, la varianza y el tamaño muestral son ingresadas por el usuario). Luego el gráfico que muestra InfoStat es construido a partir de repetidas muestras tomadas desde dicha distribución (simulación de muestreo).

El usuario puede además indicar el nivel de confianza con el que desea trabajar. Al acceder al menú aparecerá la siguiente ventana:



Al completar la información necesaria y al **Aceptar** se generará un gráfico con su correspondiente ventana de **Herramientas gráficas** como la siguiente:



En el gráfico puede observarse el centro de cada intervalo indicado por un punto y debajo del título, el porcentaje de intervalos que contienen el verdadero valor del parámetro. Nótese que hay 4 intervalos con una trama diferente, ellos son los que no contienen el parámetro. En la ventana **Herramientas gráficas** se pueden cambiar la cantidad de intervalos a mostrar (**Intervalos**), el nivel de confianza de cada intervalo construido (**Confianza**), el tamaño muestral con el que se calcula cada intervalo (**Tamaño muestral**) y el tamaño de los puntos que señalan el centro de cada intervalo (**Tamaño**). El botón **Dibujar contornos** permite unir los intervalos obtenidos a través de sus límites. Al presionar este botón aparecen los contornos y se activa la opción **Colorear** que pinta el interior de los contornos generados. Los botones **Dibujar contornos** y **Colorear** cambian su nombre una vez activados, permitiendo deshacer las acciones solicitadas (**Borrar contornos** y **Decolorar** respectivamente).

El botón **Generar otro conjunto de intervalos** permite obtener otros intervalos de confianza para la media a partir de la simulación de un nuevo conjunto de muestras aleatorias, 100 muestras en este ejemplo.

Todas las muestras posibles

Menú APLICACIONES ⇒ DIDACTICAS ⇒ TODAS LAS MUESTRAS POSIBLES, permite obtener todas las muestras posibles a partir de un conjunto de observaciones sobre una característica determinada y un conjunto de estadísticos muestrales sobre cada una de las muestras generadas. Este módulo puede ser usado con fines pedagógicos para visualización de la distribución muestral de medias, varianzas y proporciones.

InfoStat requiere la selección de una variable desde una tabla de datos, la cual puede ser de tipo entero o real. Los valores leídos son presentados en la pantalla del generador de

muestras. El usuario deberá ingresar el tamaño de las muestras a generar e InfoStat generará todas las muestras posibles, de ese tamaño, desde el conjunto de datos listado. El usuario puede seleccionar el procedimiento de muestreo aleatorio simple o sistemático (ver Estadísticas, Estimación de características poblacionales). Para cada muestra se puede requerir: la identificación de los elementos que la integran, los valores de cada observación y los estadísticos muestrales *total, media, proporción, varianza* y *varianza corregida* (está corregida por finitud, o sea multiplicada por *(N*-1*)/N).*

Al invocar el submenú **Todas las muestras posibles** aparecerá el selector de variables para indicar la columna de la tabla que contiene los valores de la variable en estudio en la población y la/s columna/s que definan particiones en caso que se desee trabajar con particiones del archivo. Al **Aceptar** aparece una ventana que permite visualizar el **Tamaño de la población**, indicar el **Tamaño muestral** y obtener el número de **Muestras posibles**. Se debe elegir el tipo de muestreo: aleatorio simple (**m.a.s**.) o **sistemático**. Al pie de la ventana aparece la **Media poblacional** y la **Varianza poblacional** de los valores de la variable indicada. En **Seleccione el resultado** se puede optar por: **Índice de los valores muestreados**, **Valores observados en la muestra**, **Media muestral**, **Total estimado en la población**, **Varianza muestral**, **Varianza muestral corregida** y **Proporción de éxitos**. Luego de definir las opciones, activar el botón **Obtener muestras**. Los valores aparecerán en un nuevo archivo de datos, que por defecto InfoStat llamará *Media Muestral* (si se deja esta opción en **Seleccione el resultado**), o bien Varianza, Proporción, etc., dependiendo del estadístico que se este muestreando.

Para ilustrar el uso de esta aplicación, considérese el siguiente procedimiento que permitirá obtener un conjunto de medias muestrales a partir de una población de 30 observaciones y visualizar la distribución de las medias:

- Crear una tabla nueva que contenga 30 filas.
- Llenar el contenido de las celdas con los valores de una variable aleatoria. En este ejemplo se utiliza la distribución normal con esperanza y varianza iguales a 100 para generar la población de partida. Esto se logra invocando el menú **Datos**, submenú **Llenar con ...,** opción **Otros**, indicando en la ventana de diálogo la distribución normal e ingresando los valores de los parámetros, para este caso media=100 y varianza=100.
- Los 30 registros obtenidos representarán a la población de la cual se obtendrán todas las muestras posibles de un tamaño dado. En este ejemplo se obtuvieron todas las muestras de tamaño 3. Esto se logra en APLICACIONES ⇒ DIDÁCTICAS ⇒ TODAS LAS MUESTRAS POSIBLES y en la ventana **Todas las muestras posibles**, en el campo **Valores en la población**, seleccione el nombre de la columna que contiene los datos poblacionales recientemente generados. Luego, en la subsiguiente ventana ingresar la información en los campos correspondientes. En este caso se ingresó tamaño muestral 3 y se activó la opción **Media muestral** para obtener las medias de cada una de las muestras generadas, presionando posteriormente el botón **Obtener muestras**. Se generará un archivo llamado Media muestral con 4060 registros.

- Para visualizar la distribución de las medias muestrales obtenidas se deberá abrir el archivo anteriormente guardado, luego ir a Menú GRÁFICOS ⇒ HISTOGRAMA. Automáticamente aparecerá un gráfico como el que se muestra a continuación:

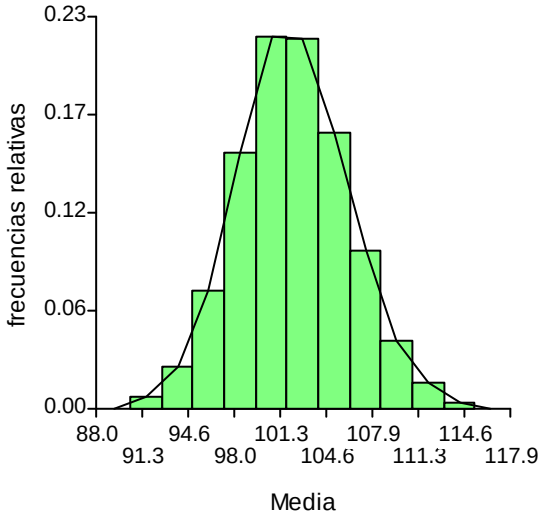
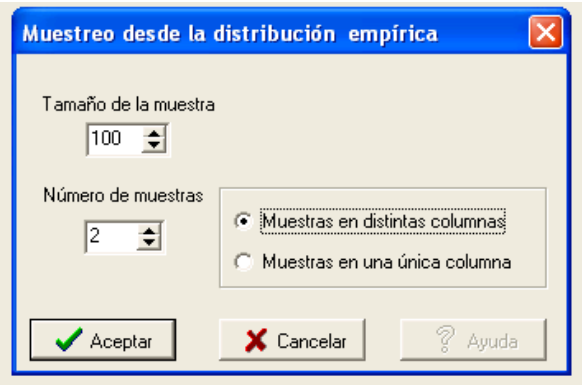


Figura 85: Histograma y polígono de frecuencias para la variable media muestral para la población de todas las muestras de tamaño 3, obtenidas por muestreo aleatorio sin reposición desde una población de 30 observaciones con distribución normal con media 100 y varianza 100.

Muestrear desde la distribución empírica

Menú APLICACIONES ⇒ DIDACTICAS ⇒ MUESTREAR DESDE LA DISTRIBUCIÓN EMPÍRICA, permite obtener muestras a partir de la distribución empírica de un conjunto de datos. El usuario debe especificar el **Tamaño de la muestra** a extraer, como así también el **Número de muestras requeridas**. Este número de muestras puede ser almacenado en una tabla de datos en varias columnas (**Muestras en distintas columnas**) o en una sola columna (**Muestras en una única columna**). Este módulo puede ser usado con fines pedagógicos para visualización de la distribución muestral de estadísticos obtenidos a partir de la distribución empírica.



Utilizando el archivo *Atriplex.idb*, se eligió la variable poder germinativo (PG) y se pidieron 2 muestras de tamaño 100 activando primeramente la opción **Muestras en distintas columnas** y luego **Muestras en una única columna**. Los archivos generados para cada caso se muestran a continuación:

Muestras

caso	PG_1	PG_2
1	15.86	20.53
2	87.58	19.95
3	34.87	18.44
4	47.12	24.62
5	80.65	85.16
6	54.00	85.53
7	30.22	59.73
8	58.87	77.46
9	63.20	64.07
10	20.79	83.17
Real	registros: 100	

Muestras

caso	Muestra	PG
1	1	15.86
2	1	87.58
3	1	34.87
4	1	47.12
5	1	80.65
6	1	54.00
7	1	30.22
8	1	58.87
9	1	63.20
10	1	20.79
Real	registros: 200	

Remuestreo

La mayor importancia de los métodos de remuestreo (muestreo desde una muestra) se refieren a situaciones donde las distribuciones de las variables aleatorias bajo estudio son desconocidas o intratables. En procedimientos inferenciales, es común la necesidad de obtener tanto una estimación del valor del parámetro distribucional de interés como el error estándar de la estimación, por ello cobran importancia los métodos generales de estimación. Importantes métodos candidatos para obtener estos valores provienen de la idea del remuestreo. *Jackknife* (Quenouille, 1949) y *bootstrap* (Efron, 1979) han probado ser métodos poderosos en numerosas situaciones ya que proveen una simulación empírica de la componente aleatoria asociado al estadístico que se esta utilizando en la estimación. La naturaleza aleatoria de la generación de muestras permite el desarrollo de una estimación empírica de la distribución en el muestreo del estadístico considerado. Estos métodos no pueden aplicarse indiscriminadamente, sino que es necesario verificar su valor en cada caso particular.

En vista de la creciente popularidad de estos métodos en diversos estudios estadísticos, se han implementado en InfoStat estos procedimientos de remuestreo básicos para facilitar la docencia sobre métodos de estimación computacionalmente intensivos. Además de poder realizar un remuestreo por jackknife o bootstrap desde una muestra original, el usuario puede guardar las muestras obtenidas y calcular estadísticos básicos sobre ellas.

Asi es posible, realizar ejercicios para comprender que para una característica poblacional θ y una única muestra de tamaño *n*, (X_1,X_2,…,X_n), es posible estimar el error estándar de un calculo muestral $\hat{\theta}$ que provee una estimación de θ.

Los pasos para estimar θ usando el procedimiento jackknife tradicional son los siguientes: (1) Dividir la muestral en n submuestras de tamaño 1, (2) Separar una submuestra desde la muestra entera original (dejar afuera un dato), (3) Calcular el estimador, **¡Error! No se pueden crear objetos modificando códigos de campo.**, desde la muestra reducida de tamaño (n-1), i.e. el estimador después de remover la i-ésima observación, i=1,2,…,n. (4) repetir los pasos 2 y 3 para todas las posibles submuestras, i.e. n veces. Estos pasos son realizados automáticamente en InfoStat al seleccionar la opcion *Jackknife* del menú de Remuestreo. El conjunto de las n estimaciones obtenidas puede ser usado para obtener un estimador puntual de θ y el error estándar de $\hat{\theta}$, seleccionando las medidas resumen apropiadas. Por ejemplo, un estimador puntual puede ser simplemente la media *jackknife* de las *n* estimaciones,

$$\hat{\theta}_{(.)} = \frac{1}{n}\sum_i \hat{\theta}_{(-i)},$$

Esta media tambien puede ser usada para desarrollar un nuevo estimador puntual, llamemoslo $\hat{\theta}_J$, mediante el calculo $\hat{\theta}_J = n\hat{\theta} - (n-1)\hat{\theta}_{(.)}$.

El estimador *jackknife* del error estándar de $\hat{\theta}$ es la raíz cuadrada de su varianza muestral calculada como: $V(\hat{\theta})_J = \frac{n-1}{n}\sum_i (\hat{\theta}_{(-i)} - \hat{\theta}(.))^2$

El coeficiente *n-1* en la ecuación de la varianza *Jackknife* es arbitrario. La racionalidad de su uso se debe al hecho de que la variación entre muestras *Jackknife* de tamaño *n-1* se espera que sea pequeña debido a que las muestras *jackknife* son n conjuntos de datos fijos obtenidos desde la muestra original. El procedimiento *bootstrap* genérico para el caso de una muestra consiste en extraer un número no fijo (como en *Jackknife*) de muestras *bootstrap*. Una muestra *bootstrap* es un conjunto de *n* observaciones de la muestra original, de tamaño n, extraidas a partir de un muestreo aleatorio con reemplazo. En una muestrea obtenida por muestreo aleatorio con reemplazo algunos elementos de la muestra original pueden aparecer más de una vez y otros no aparecer. Después de obtener automáticamente muchas muestras *bootstrap*, digamos *B* muestras, y calcular el estadistico de interés desde cada muestra *bootstrap*, $\hat{\theta}_{(i)}$, el usuario tendrá disponible un conjunto de *B* estimaciones a partir de las cuales empíricamente puede estimar θ y la varianza muestral de $\hat{\theta}$. El estimador *bootstrap* puntual del parámetro de interés es la media *bootstrap*, $\hat{\theta}_B = \frac{1}{B}\sum_i \hat{\theta}_{(i)}$

La varianza en el muestreo de $\hat{\theta}$, obtenida por *bootstrap* es la varianza muestral de los estimaciones *bootstrap*, $V(\hat{\theta})_B = \frac{1}{B-1}\sum_i (\hat{\theta}_{(i)} - \hat{\theta}_B)^2$

Como un ejemplo de aplicación, a continuación se muestran los resultados obtenidos al remuestrear un conjunto de 30 observaciones referidas al porcentaje de germinación observado en semillas de Atriplex sp. Se desean obtener estimaciones puntuales para el porcentaje promedio de germinación.

Usando menú APLICACIONES ⇒ DIDÁCTICAS⇒ REMUESTREO, se solicitó como **Tipo de muestreo** la opción ***Jackknife*** y el reporte de las **medias** muestrales (casilla activada por defecto en el panel **Guardar**). Al **Aceptar** se generó una tabla con las 30 medias muestrales. Luego, se solicitó nuevamente el remuestreo pero usando la técnica *bootstrap* para obtener 250 medias muestrales. Las tablas generadas son las siguientes:

Nueva tabla(Jackniffe)

caso	Muestra	Media
1	1.00	65.38
2	2.00	64.93
3	3.00	64.93
4	4.00	64.24
5	5.00	65.17
6	6.00	65.38
7	7.00	66.76
8	8.00	66.55
9	9.00	66.76
10	10.00	64.45
11	11.00	64.24
12	12.00	64.24
13	13.00	64.69
14	14.00	64.00
15	15.00	64.45

Real registros: 30

Nueva tabla(Bootstrap)

caso	Muestra	Media
1	1.00	65.73
2	2.00	66.03
3	3.00	69.93
4	4.00	63.07
5	5.00	62.37
6	6.00	60.73
7	7.00	66.50
8	8.00	58.53
9	9.00	73.43
10	10.00	69.33
11	11.00	55.10
12	12.00	65.63
13	13.00	68.27
14	14.00	69.63
15	15.00	57.47

Real registros: 250

Con los datos obtenidos y con los de la muestra original se calculó el valor de la media de cada conjunto de datos. Esto es, activando la tabla correspondiente, solicitar menú ESTADÍSTICAS ⇒ MEDIDAS RESUMEN y completar la información requerida. En este ejemplo la media de la variable porcentaje de germinación es 65.20, la media de las medias obtenidas por remuestreo *jackknife* es de 65.20 y la correspondiente a las medias del remuestreo *bootstrap* es 65.26.

Indices

Indices de biodiversidad

Este módulo de InfoStat ha sido diseñado y desarrollado con el asesoramiento de la Dra. Laura Pla (Universidad Nacional Experimental Francisco de Miranda, Venezuela).

Mediante este submenú, InfoStat permite obtener numerosos índices de biodiversidad. Dichos índices, como aproximación heurística para analizar comunidades vegetales y animales, se utilizan ampliamente en estudios ecológicos, de paisaje, de diversidad genética, de riesgo ambiental y de cambios en patrones de uso de las tierras.

En estudios de biodiversidad, a partir del muestreo de comunidades, el tamaño muestral (número de unidades de observación) puede ser pequeño para realizar inferencia paramétrica sobre la diversidad existente. No obstante, es deseable lograr estimaciones con niveles de confianza conocidos. Una alternativa a la estimación paramétrica para los índices de diversidad, es la construcción de intervalos de confianza (IC), mediante técnicas de computación intensiva tales como *bootstrap*.

En InfoStat se puede solicitar los siguientes índices de Biodiversidad: **riqueza** (por conteo directo), **riqueza de Chao**, índice de **Shannon-Weaver**, índice de **Simpson**, índice de **McIntosh**, índice de **Berger-Parker**, índice de **Bulla** e índice de **Kempton**. El software permite aplicar transformaciones para el cálculo de expresiones derivadas de estos índices: identidad (I), recíproca (1/I), complemento (1-I), ponderación (I/ln(riqueza)). Por ejemplo, a partir de medidas de dominancia, como el índice de Shannon, el usuario puede obtener una medida de equidad utilizando la transformación ponderación.

Los IC se obtienen por tres algoritmos: 1) *bootstrap* **estándar** o basado en la aproximación normal, 2) **percentiles** *bootstrap* y 3) *bootstrap* corregido por sesgo y aceleración (**BCA**). Aún cuando se disponga de una única muestra para estimar biodiversidad, InfoStat construirá un IC del índice seleccionado mediante la estrategia metodológica propuesta por Pla (2003).

Para estudios de diversidad regional, InfoStat permite incorporar una o más variables de clasificación de las muestras. En estos casos se obtienen estimaciones e IC de los índices para los distintos niveles de la estructura jerárquica definida por los criterios de clasificación y para el total de las muestras.

El análisis cuantitativo de la diversidad a través del cálculo de índices puede realizarse de dos formas básicas:

a) Cuando existe una única muestra por comunidad
b) Cuando existen dos o más muestras para una misma comunidad

a) Biodiversidad basada en una única muestra

En el caso de una única muestra los datos que serán procesados por InfoStat suponen una medida de abundancia (absoluta o relativa) de cada especie en la muestra. La tabla de datos debe contener una columna para cada especie y en la fila (caso) las abundancias de cada especie en la muestra.

b) Biodiversidad basada en dos o más muestras

En el caso de varias muestras para una misma comunidad los datos que serán procesados por InfoStat suponen una medida de abundancia (absoluta o relativa) de cada especie (variables o columnas de la tabla de datos) en cada muestra o caso (una transecta, una subtransecta, una celda de una cuadrícula). Adicionalmente la tabla de datos puede contener una o más columnas con variables de clasificación, por ejemplo comunidad, localidad, región. Siempre que la tabla de datos contenga varios casos como subunidades de una misma muestra debe incluirse como criterio de clasificación una columna que identifique cada muestra.

En el Menú APLICACIONES ⇒ OTROS ⇒ ÍNDICES DE DIVERSIDAD, aparecerá la ventana **Medidas de diversidad**, donde el usuario debe indicar en el panel **Variables** las especies, o unidades que componen la biodiversidad, y en **Criterios de clasificación** identificar las variables del archivo que permiten diferenciar muestras (por ejemplo transectas, campo, región). InfoStat calculará índices de biodiversidad y sus intervalos de confianza para cada muestra, para cada nivel de la variable de clasificación y para el total. Si hay más de un criterio de clasificación la primera variable debe corresponder a la jerarquía principal y la siguiente variable se anida en la predecesora. InfoStat acepta más de dos variables de clasificación.

Al **Aceptar**, se muestra una ventana, donde se debe seleccionar el/los índices a estimar. En el panel **Intervalos Bootstrap** seleccione el método de estimación *bootstrap*, con el cual se obtendrán las medidas de confianza. En la misma ventana de diálogo indique el coeficiente de confianza para el IC a obtener (**Confianza%**), el número de **muestras *bootstrap*** con los que se realizan las estimaciones y alguna de las siguientes **transformaciones** del índice: **identidad** (I), recíproca (**1/I**), complemento (**1-I**) o ponderada (**I/ln(r)** donde r es riqueza).

Si no se especifican variables de clasificación, InfoStat calcula los índices e IC para cada caso o muestra. Si se especifican variables de clasificación y se desean también los IC para cada caso debe activarse la casilla **Estimaciones bootstrap para casos** e incorporar la variable casos en la subventana **Criterios de clasificación**.

Si se activa la casilla **Totales por nivel**, InfoStat incorpora a la ventana resultados las frecuencias totales de cada nivel de las variables de clasificación para las cuales se calculó el índice.

Breve descripción de los Métodos implementados para cuantificara biodiversidad.

Si partimos de una población con un número total S de clases (típicamente especies en estudios de biodiversidad) no superpuestas, identificadas por i=1, 2, ... , S. Designamos por π_i la proporción de la i-ésima clase. Como las clases son excluyentes, es decir que un mismo elemento o individuo no puede pertenecer a más de una clase, los π_i están sujetos a la restricción que $\sum_{i=1}^{S} \pi_i = 1$.

Suponga que se toma una muestra aleatoria de esta población, llamaremos X_i al número de individuos o abundancia de la clase i; si $X_i=0$, entonces la i-ésima especie no se ha

observado en la muestra. Al número total de especies observadas en la muestra la llamamos r, que nunca podrá ser mayor que S.

Denominaremos f_k al número de clases con frecuencia o abundancia k y así, el número total de individuos o abundancia total (to) en la muestra puede calcularse como:

$$to = \sum_{i=1}^{r} X_i$$

que puede ser expresado en función de los f_k como: $to = \sum_{\forall k} k f_k$

Asimismo la estimación de los π_i desde una muestra puede hacerse a partir de las frecuencias relativas como:

$$p_i = \frac{X_i}{to}$$

Los cálculos de riqueza se basan en las frecuencias de simples y dobles (f_1 y f_2) y en el número de clases o especies efectivamente observadas (r). Los cálculos de los índices de biodiversidad se basan en las p_i o frecuencias relativas de cada especie.

Riqueza

InfoStat puede calcular la riqueza observada en la muestra (r), que es el número total de especies presentes en esa muestra. Esta estimación es siempre un valor no mayor a la verdadera riqueza de la comunidad.

Riqueza Chao: Chao (1987) derivó un estimador para el número total de especies presentes en una comunidad como:

$$S = r + \frac{f_1^2}{2f_2}$$

siendo f_1 el número de especies con abundancia unitaria y f_2 el número de especies con abundancia dos. Cuando no existen especies con abundancia 2 en la muestra, el índice no puede calcularse.

Índice de Shannon

El índice de biodiversidad de Shannon (Shannon y Weaver, 1949) se basa en suponer que la heterogeneidad depende del número de especies presentes y de su abundancia relativa. Es una medida del grado de incertidumbre asociada a la selección aleatoria de un individuo en la comunidad. Se calcula como:

$$H = -\sum_{i=1}^{r} p_i \ln p_i$$

La diversidad máxima (H=ln *r*) se alcanza cuando todas las especies están igualmente presentes. Un índice de equidad asociado a esta medida de diversidad puede calcularse como el cociente $H/H_{max}=H/\ln r$.

Con presencia de varias observaciones por muestra o estructuras jerárquicas de clasificación se recomienda la construcción de IC por el método BCA (Pla, 2001). Para muestras únicas con frecuencias o abundancias totales de hasta 500, se recomienda el uso del método ajustado para la estimación del índice de Shannon H^a como:

$$H^a = 2.73\,H - 1.75\,H^* + 0.0003\,r^*$$

siendo H el índice calculado con la ecuación (1.4), H^* la estimación bootstrap del índice y r^* la estimación bootstrap de la riqueza. La construcción de los IC se basa en la desviación estándar bootstrap y sus límites se calculan como:

$$H^a \pm z_{\alpha/2}\, s^*_r$$

siendo s^*_r la desviación estándar bootstrap de la riqueza (Pla, 2004). Para frecuencias o abundancias totales en la muestra mayores de 500 se recomienda directamente el uso de IC por BCA.

Índice de Simpson

Propuesto por Simpson (1949) sugiere que una medida intuitiva de la diversidad de una población está dada por la probabilidad de que dos individuos tomados independientemente de una población pertenezcan a la misma especie.

El estimador del índice de Simpson se calcula como:

$$D = \sum_{i=1}^{r} \frac{x_i(x_i - 1)}{to(to - 1)}$$

El índice varía entre $1/r$ (menor concentración o máxima diversidad posible con r especies) y uno (mayor concentración o mínima dispersión cuando una especie domina la comunidad). El recíproco de Simpson ($1/D$) puede interpretarse como el número de especies igualmente abundantes necesarias para producir la heterogeneidad observada en la muestra. No se recomienda la construcción de intervalos de confianza cuando D es menor que 0.02 o su recíproco es mayor que 50 (Pla, 2003).

Índice de McIntosh

El índice de McIntosh (1967) es un índice de dominancia que se basa en considerar que la comunidad es un punto en un hiperespacio definido por las especies, y que puede cuantificarse como la distancia euclidiana desde ese punto al origen ($\sqrt{\sum_{i=1}^{r} x_i^2}$). Si hay tantas especies como individuos la diversidad es máxima, la diferencia entre este máximo (to) y la comunidad en estudio es una medida de la diversidad absoluta (numerador del índice).

El estimador del índice se calcula como:

$$U = \frac{to - \sqrt{\sum_{i=1}^{r} x_i^2}}{to - \sqrt{to}}$$

y puede interpretarse como la proporción de la diversidad máxima absoluta (el denominador de la ecuación 1.8) para esa frecuencia total. Varía entre cero (mínima diversidad, cuando hay una sola especie) y uno (cuando la diversidad es máxima y todas las especies tienen frecuencia uno).

Índice de Bulla

En un gráfico de la frecuencia relativa de aparición de las especies (ordenadas) *versus* el número de especies (abscisas) una línea horizontal en `1/r` representaría una comunidad con diversidad máxima. Si se superpone a ésta una línea que representa la frecuencia relativa en la comunidad y se calcula el grado de solapamiento entre estas dos distribuciones $\left(\sum_{i=1}^{r} \min\left(p_i, \frac{1}{r}\right)\right)$ se obtiene la medida de equidad que propone Bulla (1994).

El índice se calcula como:

$$O = \frac{\left[\sum_{1=1}^{r} \min(p_i, \frac{1}{r})\right] r - 1}{r - 1}$$

luego se ajusta para que varíe entre `cero`, cuando una especie aparece con dominancia absoluta, y `uno` cuando todas las especies están igualmente presentes.

Índice de Berger-Parker

Propuesto originalmente para poblaciones de fitoplancton (Berger y Parker, 1970), toma en cuenta sólo la especie más abundante y es el más simple de los índices de biodiversidad. Se calcula como:

$$d = x_{max} / to.$$

Índice de Kempton

Propone un índice que evita el excesivo peso de las especies más abundantes y de las especies menos abundantes en el cálculo de la diversidad. El índice (Kempton y Taylor, 1976) es la pendiente entre el primer cuartil (p=0.25) y el tercer cuartil (p=0.75) de la distribución logarítmica acumulada de la abundancia de las especies en orden descendente (abscisas) y las especies (ordenadas). InfoStat lo calcula como:

$$Q = \frac{\frac{1}{2} f_{k=.25r} + \frac{1}{2} r + \frac{1}{2} f_{k=.75r}}{\log\left(\sum_{k=1}^{k=.75r} k f_k \; / \; \sum_{k=1}^{k=.25r} k f_k \right)}$$

El índice tiende a cero cuando el primer y tercer cuartil coinciden, o sea que las especies del 'centro' de la distribución aportan muy poco a la abundancia acumulada. Cuanto más homogénea es la distribución de la abundancia entre las especies mayor será el índice de Kempton. Los IC calculados por cualquiera de los métodos bootstrap no poseen buen comportamiento cuando se trata de una única muestra, y no existe otro método para calcular la desviación estándar cuando se desconoce el modelo probabilístico de distribución de origen.

Intervalos bootstrap calculados por InfoStat

Un intervalo de confianza de nivel α se define como un conjunto de valores del parámetro (intervalo) que con confianza $(1-\alpha)100\%$ incluirían el valor del parámetro en la población, dada la variabilidad y la distribución muestral del estimador en la muestra observada. Así, los intervalos de confianza paramétricos se construyen a partir de suposiciones sobre la forma de la distribución muestral del estimador (Normal, t de Student, Chi cuadrado, etc.). En el caso de los índices de biodiversidad no es razonable suponer una distribución muestral del estimador y por eso InfoStat permite utilizar una técnica de construcción de intervalos no paramétrica basada en el procedimiento de remuestreo conocido como *bootstrap*.

La técnica de *bootstrap* consiste en extraer al azar mediante un muestreo con reposición *B* muestras de tamaño *n* desde la muestra original de tamaño *n*. En cada una de las *B* muestras bootstrap (por defecto *B*=500) InfoStat calculará el estadístico de interés (en este caso un índice de biodiversidad).

Cuando se selecciona una variable de clasificación y se calcula el índice de biodiversidad para un conjunto de filas, *n* corresponde al número total de filas utilizadas para esa estimación. InfoStat totalizará por especie y calculará el índice solicitado en cada muestra bootstrap.

Para aplicar métodos de remuestreo al cálculo del índice para una única muestra InfoStat supone que la población de la cuál proviene la muestra es homogénea y dividida en 'porciones unitarias' cada una tan grande como el mínimo tamaño reconocible, expresado como frecuencia. Así, la abundancia de cada especie se subdivide en x_i porciones unitarias que pueden ser muestreadas. Como la abundancia se expresa en la misma escala para todas las especies, tanto en la población como en las muestras (y las remuestreas) podemos hacer una selección aleatoria de 'porciones unitarias' caracterizadas por la especie que se registra en ella. Estos son *N* objetos virtuales, que se suponen mutuamente independientes y que tienen la misma distribución (no conocida) de probabilidad. Este valor de *N* es el que InfoStat toma en estos casos como 'tamaño de la muestra'.

Intervalo por percentiles

Es posible ordenar ascendentemente las *B* estimaciones e identificar los cuantiles que serán utilizados como límites del intervalo de confianza bootstrap del parámetro de interés. Así, seleccionando **Estimación por percentiles** los límites del intervalo bilateral con confianza (1-α)×100, corresponden a los percentiles (α/2)×100 y (1-α/2)×100 de la lista de estimaciones obtenidas en *B* muestras bootstrap extraídas de la muestra original.

Intervalo estándar por aproximación normal

Se basa en suponer que el estimador bootstrap I* tiene una distribución aproximadamente normal con media μ y desviación estándar σ; y que por lo tanto hay una probabilidad de (1 - ∀) que

$$I + z_{\alpha/2}\sigma_I < I^* < I + z_{1-\alpha/2}\sigma_I$$

para toda muestra aleatoria usada en la estimación (Efron 1979).

Sobre esta base es posible estimar los límites del intervalo de confianza como

$$LI = \bar{I}^* + z_{\alpha/2}s_I^*$$

$$LS = \bar{I}^* + z_{1-\alpha/2}s_I^*$$

siendo $\bar{I}^*$ el estimador boostrap del índice, s_I^* es estimador bootstrap de la desviación estándar y $z_{\alpha/2}$ y $z_{1-\alpha/2}$ los correspondientes valores de la distribución normal estándar.

Intervalo corregido por sesgo y aceleración (BCA)

Para aplicar este método se requiere

- La distribución bootstrap del estimador I*
- Una estimación de la mediana del sesgo de la estimación, es decir cuál es la diferencia entre la media (o la mediana) de la distribución muestral de I y θ (el parámetro deseado).
- Una estimación de la aceleración de la varianza, es decir cuánto aumenta o disminuye la varianza muestral a medida que θ (el parámetro) se incrementa.
- El valor de α, para determinar el nivel de confianza del intervalo (1 - 2α)100%

El cálculo de la aceleración (a) es realiza en forma iterativa como:

$$a = \frac{\sum_{i=1}^{n} (I_{(-i)}^* - \bar{I}_{(-i)}^*)^3}{6 * (\sum_{i=1}^{n} (I_{(i)}^* - \bar{I}_{(-i)}^*)^2)^{3/2}}$$

siendo $I^*_{(-i)}$ el estimador del parámetro en una muestra reducida en la cuál se ha omitido la i-ésima observación; $\overline{I}^*_{(-i)}$ la media de los estimadores calculados con la muestra reducida; y n el número total de muestras.

Para el cálculo de los límites se utiliza el método del percentil corregido (Efron 1981). Esta técnica remueve el sesgo que resulta del fracaso de la mediana en estimar la distribución del estimador del parámetro de interés. Se puede hacer usando las muestras bootstrap para representar la verdadera distribución bootstrap, y suponiendo que existe una transformación monotónica que se distribuye como una normal de media cero y varianza **uno**.

Supongamos que de 1000 estimaciones bootstrap se obtienen 550 mayores que la muestra original. Entonces, el cociente 550/1000 = 0.55 (llamado p*) debe reemplazar el percentil de la mediana. Conocer este cociente mejora la estimación de los límites del intervalo de confianza y rescata la posible asimetría de la distribución.

En el cálculo del IC se utilizan los valores de la distribución normal estándar $z_{(\forall/2)}$, $z_{(1 - \forall/2)}$, y z_{p^*} que corresponden a las probabilidades $\forall/2$, $(1-\forall/2)$ y p^*, respectivamente. Si p^* is 0.50 entonces $z_{p^*} = 0$, y se obtendrá un intervalo simétrico.

El intervalo de confianza combinando la corrección por sesgo y por aceleración corresponderá a los cuantiles de la distribución empírica que coincidan con:

$$q_{LI} = z_{|p^*|} - \frac{z_{|1-\alpha|} - z_{|p^*|}}{1 + a\left(z_{|1-\alpha|} - z_{|p^*|}\right)}$$

para el límite inferior, y con:

$$q_{LS} = z_{|p^*|} + \frac{z_{|p^*|} - z_{|\alpha|}}{1 - a\left(z_{|p^*|} - z_{|\alpha|}\right)}$$

para el límite superior.

Bibliografía

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons, Inc., New York.

Altman D.G. (1991). *Practical Statistics for Medical Research*. Chapman and Hall. London.

Alvarez, L.J., Delrieu, J.C. y Jareño, J. (1993). Tratamiento de predicciones conflictivas: Empleo eficiente de información extramuestral. *Estadística Española*, 35: 439-461.

Anderberg, M.R. (1973). *Cluster analysis for applications*. New York. Academic Press.

Anderson, V.L. y Mc Lean, R.A. (1974). *Design of Experiments: A Realistic Approach*. Marcel Dekker Inc., New York.

Bautista, M.G., Smith, D.W. y Steiner, R.L. (1997). A cluster-based approach to mean separation. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(2): 179-197.

Berger, W.H.; Parker, F.L. (1970). Diversity of planktonic foraminifera in deep-sea sediments. *Science*, 168: 1345-1347.

Box, G.E.P. (1953). Non-normality and Tests on Variance. *Biometrika*, 40: 318-335.

Box, G.E.P. y Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Bramardi, S. (2001). Minicurso sobre métodos estadísticos multivariados para la caracterización de recursos fitogenéticos. VI Reunión Científica del Grupo Argentino de Biometría.

Bulla, L. (1994). An index of evenness and its associated diversity measure. *Oikos,* 70: 167-171.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics,* 43: 783-791.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Association*, 74: 829-836.

Cochran, W.G. y Cox, G.M. (1957). *Experimental Designs*. John Wiley & Sons, Inc. London.

Cole, J.W.L.; Grizzle, J.E. (1966). Applications of Multivariate Analysis of Variance to Repeated Measures Experiments. *Biometrics*, 22: 810-828.

Collings, B.J.; Hamilton, M.A. (1988). Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics*, 44: 847-860.

Conover, W.J. (1999). *Practical Nonparametric Statistics*. John Wiley & Sons, Inc., New York.

Di Rienzo, J.A.; Guzmán A.W.; Casanoves F. (2002). A Multiple Comparisons Method based on the Distribution of the Root Node Distance of a Binary Tree. *Journal of Agricultural, Biological, and Environment Statistics*, 7(2): 1-14.

Di Rienzo, J; Casanoves, F.; Gonzalez, L.; Tablada, E.; Díaz M.; Robledo, C.; Balzarini, M. (2001). *Estadística para las Ciencias Agropecuarias*. 4ta Ed. Triunfar. Córdoba, Argentina.

Draper, N.R. ; Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons Inc., New York, 3rd Ed.

Duncan, A. J. (1974). *Quality Control and Industrial Statistics*. 4th Ed., Irwin, Homewoods, III.

Durbin, J. (1960). Estimation of Parameters in Time-Series Regression Models. *Journal of the Royal Statistical Society,* Series B, 22: 139-153.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7: 1-26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9: 139-172.

Efron, B; Tibshirani, R. (1993). Bootstrap methods for standard errors, confidence intervals, and other methods of statistical accuracy. *Statistical Science*, 1(1): 54-77.

Einot, I.; Gabriel, K.R. (1975). A Study of the Powers of Several Methods of Multiple Comparisons. *Journal of the American Statistical Association*, 70: 574-583.

Everitt, B. (1974). *Cluster analysis*. London. Heinemann Educational Books.

Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomy Problems. *Annals of Eugenics*, 7: 179-188.

Florek, K., Lukaszewicz, J., Perkal, J. y Zuvrzycki, S. (1951a). Sur la Liaison et la division des Points d´un Ensemble Fini. *Colloquium Mathematicae,* 2: 282-285.

Florek, K., Lukaszewicz, J., Perkal, J. y Zuvrzycki, S. (1951b). Taksonomia Wroclawska. *Przeglad Antropol*., 17: 193-211.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of American Statistical Association*, 32: 675-701.

Gabriel, K.R. (1971). Biplot display of multivariate matrices with application to principal components analysis. *Biometrika,* 58: 453-467.

Gabriel, K.R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In V. Barnett (Ed.), *Interpreting Multivariate Data*. London: Wiley.

Gonzalez, L.A. (2001). *Extensión no paramétrica de una técnica de comparaciones múltiples basada en la distribución del nodo raíz de un árbol binario*. Tesis de Maestría. Magister en Estadística Aplicada, Universidad Nacional de Córdoba.

Gower, J. C. (1975). Generalizad procrustes analysis. Psychometrika, 40:33-51.

Gower, J.C., Ross, P.G.N. (1969). Minimun spanning tress and single cluster analysis. *Applied Statistics*, 18:54-64.

Gower, J.C.; Digby, P.G.N. (1981). Expressing complex relationships in two dimensions, in Interpreting Multivariate Data (V. Barnett, ed.), pp. 83-118, John Wiley & Son, Inc.

Graybill, F.A. (1961). *An Introduction to Linear Statistical Models*. McGraw Hill, New York.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M.J. (1988). Correspondence Analysis of Multivariate Categorical Data by Weighted Least-Squares. *Biometrika*, 75: 457-467.

Greenacre, M.J. (1994). *Multiple and Joint Correspondence Analysis*. En Greenacre, M.J. y Blasius, J. (Ed.). *Correspondence Analysis in the Social Sciences*, London: Academic Press.

Greenacre, M.J.; Hastie, T. (1987). The Geometric Interpretation of Correspondence Analysis. *Journal of the American Statistical Association*, 82: 437-447.

Greenhouse, S.W.; Geisser, S. (1959). On Methods in the Analysis of Profile Data. *Psychometrika*, 32: 95-112.

Guerrero, V.M.; Peña, D. (2000). Linear Combination of information in Time Series with Linear Restrictions. *Journal of Forecasting*, 19: 103-122.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton, New Jersey: Princeton, Univ. Press.

Hartigan, J.A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76: 388-394.

Hinkelmann, K.; Kempthorne, O. (1994). *Design and Analysis of Experiments*. Volume 1: Introduction to Experimental Design. John Wiley & Sons, Inc., New York.

Hocking, R.R. (1996). *Methods and Applications of Linear Models*. Regression and the Analysis of Variance. John Wiley & Sons, Inc., New York.

Hollander, M.; Wolfe, D.A. (1999). *Nonparametric Statistical Methods*. John Wiley & Sons, Inc., New York.

Hosmer, D.W.; Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons, Inc., New York.

Hotelling, H. (1936). Relations Between Two Sets of Variables. *Biometrika*, 28: 321-377.

Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*. London: Chapman & Hall.

Hsu, J.C.; Nelson, B. (1998). Multiple Comparisons in the General Linear Model. *Journal of Computational and Graphical Statistics*, 7(1): 23-41.

Huynh, H.; Feldt, L.S. (1970). Conditions under Which Mean Square Ratios in Repeated Measurements Designs have Exact F-Distributions. *Journal of the American Statistical Association*, 65: 1582-1589.

Huynh, H.; Feldt, L.S. (1976). Estimation of the Box Correction for Degrees of Freedom from Sample Data in the Randomized Block and Split Plot Designs. *Journal of Educational Statistics*, 1: 69-82.

Jarque, C.M;. Bera, A.K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55: 163-172.

Johnson, R.A.; Wichern, D.W. (1998). *Applied multivariate statistical analysis*. 4th Ed. Prentice Hall, Upper Saddle River, NJ.

Jolliffe, I.T. (1975). Cluster analysis as a multiple comparison method. *Applied Statistics* (North-Holland, Amsterdam), 159-168.

Kempton, R.A.; Taylor, L.R. (1976). Models and statistics for species diversity. *Nature*, 275: 818-820.

Kruskal, W.H.; Wallis, W.A. (1952). Use of ranks on one-criterion variance analysis. *Journal of the American Statistical Association,* 47: 583-621.

Lebart, L., Morineau, A.; Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: John Wiley & Sons, Inc.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, Inc. San Francisco. USA.

Mahibbur, R.M.; Govindarajulu, Z. (1997). A modification of the test of Shapiro and Wilks for normality. *Journal of Applied Statistics*, 24(2): 219-235.

Mantel, N. A. (1967). The detection of disease clustering and a generalized regression approach. Cancer Res., 27:209-220.

Marascuilo, L.; McSweeney, M. (1977). *Nonparametric and Distribution-Free Methods for the Social Sciences*. Wadsworth Publishing Company, Inc. U.S.A.

McIntosh, R.P. (1967). An index of diversity and the relation of certain concepts to diversity. *Ecology* 48: 392-404.

McQuitty, L.L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurements*, 26: 825-831|.

Miller, R.G. (1981). *Simultaneous Statistical Inference*. 2nd Ed. Springer-Verlag, Heidelberg and Berlin.

Milligan, G.W. (1980). An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika*, 45: 325-342.

Montgomery, D.C. (1991). *Control Estadístico de la Calidad*. Grupo Editorial Iberoamérica.

Montgomery, D.C. (1991). *Diseño y Análisis de Experimentos.* Grupo Editorial Iberoamérica.

Morrison, D.F. (1976). *Multivariate Statistical Methods*. 2nd Ed., New York: McGraw-Hill Book Co.

Nelder J.A.; Mead R. (1965). Downhill simplex method in multidimensions. *Computer Journal,* 7: 308-315.

Nelder, J.A. (1994). The Statistics of Linear Models: Back to Basics. *Statistics and Computing*, 4: 243-256.

Ostle, B. (1977). *Estadística Aplicada. Técnicas de la Estadística Moderna, cuando y donde aplicarlas*. Editorial Limusa, México.

Pillai, K.C.S. (1960). *Statistical Tables for Tests of Multivariate Hypotheses. Manila*, The Statistical Center, University of the Philippines.

Pindyck, R.S.; Rubinfeld, D.L. (1999). *Econometría: Modelos y pronósticos*. McGraw-Hill. 4th Ed.

Pla, L. (2003). Confidence intervals for α-biodiversity indices. *II Reunión de Biometría de la Región Centroamérica y Caribe de la Sociedad Internacional de Biometría (RCAC-IBS).* Universidad de Puerto Rico, Mayagüez.

Pla, L. (2004). Bootstrap confidence interval for Shannon biodiversity index: a simulation study. *Journal of Agricultural, Biological and Environmental Statistics*. Aceptado para su publicación en febrero 2003.

Pla, L.; Matteucci, S. (2001). Intervalos de confianza bootstrap del índice de biodiversidad de Shannon. *Revista de la Facultad de Agronomía LUZ* 18: 222-234.

Press, W.H., Flannery, P.; Vetterling W.T. (1986). *Numerical Recipes*. Cambridge University Press.

Quenouille, M H. (1949). Aproximate tests of correlation in time-series. *Journal of the Royal Statistical Society.* Series B, 11: 68-84.

Rawlings. J.O. (1988). *Applied Regression Analysis: a Research Tool*. Wadsworth & Brooks/Cole Advance Books & Software.

Ryan, T.P. (1997). *Modern regression methods*. JohnWiley & Sons Inc., New York.

Schöemann, P. H. y Carrol, R. M.. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, 35:245.

Scott, A. J.; Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics,* 30: 507-512.

Searle, S.R. (1971). *Linear Models*. New York, John Wiley & Sons, Inc., New York

Searle, S.R. (1987). *Linear Models for Unbalanced Data*. New York, John Wiley & Sons, Inc., New York.

Seber, G.A.F. ;Wild, C.J. (1989). *Nonlinar Regresión*. New York, John Wiley & Sons Inc., New York.

Shannon, C.; Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press, Urbana-USA.

Shapiro, S.S.; Francia, R.S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association,* 67: 215-216.

Simpson, E.H. (1949). Measurements of diversity. *Nature* 163: 688.

Smrekar, M.R., Robledo, C.W., Di Rienzo, J.A. (2001). Predicción de valores perdidos con regresión restringida en modelos ARIMA: evaluación Monte Carlo bajo modelos mal especificados. *XXIX Coloquio Argentino de Estadística*. Argentina.

Snedecor, G.W. (1956). *Métodos Estadísticos Aplicados a la Investigación Agrícola y Biológica*. Compañía Editorial Continental, S.A., México.

Snedecor, G.W.; Cochran, W.G. (1967). *Statistical Methods*. Ames, IA: Iowa State University Press.

Sokal, R.R.; Michener, C.D. (1958). A Statistical Methods for Evaluating Systematic Relationships. *University of Kansas Science Bulletin*, 38: 1409-1438.

Sorensen, T. (1948). A Method of Establishing Groups of Equal Amplitude in Plant Sociology based on Similarity of Species Content and its Application to analyses of Vegetation on Danish Commons. *Biologiske Skrifteer*, 5: 1-34.

Spath, H. (1980). *Cluster Analysis Algorithms*. Chichester, England: Ellis Horwood.

Timm, N.H. (1975). *Multivariate Analysis*. Brooks-Cole Publishing Co., Monterrey, CA

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58: 236-244.

Winer, B.J. (1971). *Statistical Principles in Experimental Design*. 2nd Ed., McGraw-Hill Book Co., New York.

Indice de contenidos

D