

Manual de prácticas para la asignatura Análisis de Datos en Ingeniería Informática

Jorge Mateu
Departamento de Matemáticas, Universitat Jaume I

1 PRACTICA 1. INTRODUCCION

Procedimientos generales para el uso del programa Statgraphics

PROCEDIMIENTOS GENERALES

La versión Windows del Statgraphics muestra una ventana (Statfolio) con una barra de opciones dentro de las cuales encontramos varios submenús agrupados en bloques temáticos:

File Permite la creación y manipulación de archivos. Entre las opciones destaca: Open (Abre), Close (Cierra) Statfolio, un archivo de datos o una Statgallery.

Edit Permite labores relacionadas con la edición de archivos.

Plot Permite acceder a los diferentes tipos de gráficos disponibles en el programa.

Describe Permite un análisis descriptivo de los datos.

Compare Permite acceder a tipos de análisis que generan y comparan estadísticos descriptivos.

Relate Permite acceder a análisis que modelizan la relación entre variables dependientes e independientes.

Special Módulos más especializados y relacionados con técnicas estadísticas avanzadas.

View Visualización de opciones de trabajo.

Window Visualización de ventanas de trabajo.

Help Menú de ayuda.

USO DE LOS DIFERENTES TIPOS DE VENTANAS

Una ventana puede representar muchos objetos. En este software encontramos diferentes tipos de ventanas:

Ventana de aplicación (Application window) Aparece cuando entramos por primera vez en el programa. Contiene (considerando la pantalla desde el extremo superior al inferior) la barra de título, la barra de menús, la barra de herramientas de aplicación y en la parte inferior, el botón de la ventana de los comentarios (The untitled comment bar), el botón de la ventana de datos (the untitled button), el Statadvisor button y la StatGallery button.

Ventana de comentarios (Comments window) Aparece cuando se selecciona 'untitled Comments taskbar button'. Sirve para recordar información sobre los datos y se puede usar como un cuaderno de notas (funciona como un editor de textos pudiéndose copiar, cortar, pegar...).

Ventana de cálculo (Spreadsheet window) Aparece cuando se selecciona 'untitled Data taskbar button'. Esta ventana se usa para crear nuevas hojas de cálculo, modificar las ya existentes o modificar datos usando el editor. La usaremos para crear nuestros archivos de datos.

Ventana de análisis (Analysis window) Aparece después de que hallamos seleccionado un gráfico o un análisis estadístico de la barra de menús. Proporciona la información correspondiente referida al dibujo o una pantalla de diálogo para el análisis.

La ventana de análisis contiene tres componentes:

1. Analysis Icon Title Bar: Muestra el icono para el análisis y su título.
2. Analysis Toolbar: Muestra los botones que se pueden usar para trabajar en la ventana de diálogo, acceder a las opciones gráficas y de tablas, salvar los resultados de ciertos análisis y cuando sea posible acceder a opciones gráficas adicionales.
3. Text and graphics panes: Muestran el texto y los gráficos del análisis en uno o dos paneles.

Ventana StatGallery (StatGallery window) Aparece cuando seleccionamos el botón 'StatGallery Icon'. Sirve para arreglar y ordenar textos y gráficos que se quieren mostrar o imprimir. Nos proporciona información sobre el análisis.

Ventana de presentación preliminar (Preview window) Aparece cuando seleccionamos el submenú 'Print Preview' del menú File.

CREACIÓN Y MANEJO DE FICHEROS DE DATOS

- *Creación de un fichero y manejo de las variables contenidas en él.*

La manera más cómoda de trabajar con Statgraphics es tener los datos guardados en ficheros de datos. Para crear un archivo de datos:

1. Seleccionamos en la barra de tareas principal, dentro de la opción **Window** el archivo **untitled**.
2. Seleccionamos con el botón izquierdo del ratón **Col1**, el nombre por defecto de la variable.
3. Con el botón de la derecha del ratón seleccionamos el menú que aparece la opción **Modify Column** y rellenamos los campos que aparecen:

- (a) Escribimos el nombre de la variable (Name).
- (b) Podemos añadir comentarios sobre la variable (Comment).
- (c) Especificamos la anchura que queremos para la columna (Width).
- (d) Especificamos el tipo de variable: *Numeric* (variables numericas con tantos decimales como queramos); *Character* (variables cualitativas o números tratados como valores cualitativos); *Integer* (valores enteros); *Fixed Decimal* (números con una cantidad de decimales prefijada); *Formula* (permite generar variables mediante fórmulas a partir de otras existentes).
- (e) Volvemos a la columna de la variable que queremos generar y vamos introduciendo los valores.

4. Repetimos el proceso para cada una de las variables que queramos introducir y para guardar los datos seleccionamos dentro de la opción FILE, el submenú SAVE AS y SAVE DATA FILE AS.

- *Manejo de Ficheros. File Utilities.*

Las operaciones con ficheros de datos las podemos hacer con la opción FILE.

- *Guardar los resultados y recuperarlos.*

Se hace fácilmente con el menú FILE (con OPEN, SAVE y PRINT).

MANEJO DE DATOS Y DESCRIPCIONES UNIVARIANTES

Tablas de frecuencias

Usamos **Describe** y seleccionamos **Categorical data** y posteriormente **Tabulation**. En el campo **Data** escribimos el nombre de la variable y pulsamos en **OK**.

Aparecerá la ventana de **Tabulation** con una barra de botones en la que pulsaremos el segundo por la izquierda.

Como consecuencia aparecerá una nueva ventana **Tabular option** en la que marcaremos el recuadro correspondiente a **Frequency table**.

Histogramas y Polígonos de frecuencias

Empleamos **Plot** seleccionando **Exploratory plots** y como opción **Frequency histogram** (existe la posibilidad de ir directamente si está cargada la **Toolbar** a través del botón que tiene el dibujo de un histograma, exactamente en el centro).

Si queremos cambiar los parámetros del histograma o realizar un polígono de frecuencias, pulsaremos dos veces sobre el gráfico y cuando se maximice pulsaremos sobre cualquiera de los ejes. A continuación debemos pulsar esta vez el botón derecho del ratón y seleccionar **Pane options** eligiendo como **Lower limit** el extremo inferior del primer intervalo, **Upper limit** será el extremo superior del último intervalo y **Number of classes** el número de intervalos. También podemos marcar **Relative** o **Cumulative** si queremos que las frecuencias del gráfico sean relativas, acumuladas o ambas cosas. Finalmente en cuanto a **Plot Type** podemos seleccionar **Histogram** o **Polygon** según sea nuestro objetivo.

Diagramas de Sectores

Usando **Describe**, seleccionando **Categorical data** y posteriormente **Tabulation**. En el campo **Data** escribimos el nombre de la variable y pulsamos **OK**.

Aparecerá la ventana de **Tabulation** con una barra de botones en la que pulsaremos el tercero por la izquierda seleccionando **Piechart**.

Si se quiere cambiar los parámetros se deberá pulsar dos veces sobre el gráfico y una vez ampliado pulsar en el interior del círculo con el habitual botón izquierdo del ratón y después con el derecho para finalmente seleccionar **Pane Options**.

En el rectángulo **Legende** elegiremos la opción para la leyenda que aparecerá en el margen superior izquierdo: porcentajes (**percents**), frecuencias absolutas (**counts**), nombre de las categorías (**labels**) o nada (**none**). Las mismas opciones se permiten para el rectángulo **Labels** que se refiere al dato que aparece al lado de cada sector.

Diagramas de Barras

Usando **Describe**, seleccionando **Categorical data** y posteriormente **Tabulation**. En el campo **Data** escribimos el nombre de la variable y pulsamos **OK**.

Aparecerá la ventana de **Tabulation** con una barra de botones en la que pulsaremos el tercero por la izquierda seleccionando **Barchart**.

Si se quiere cambiar los parámetros se deberá pulsar dos veces sobre el gráfico y una vez ampliado pulsar en el interior del círculo con el habitual botón izquierdo del ratón y después con el derecho para finalmente seleccionar **Pane Options**.

Las **Pane Options** son, en este caso, **Clustered** (normal) o **Stacked** (acumulado), **Frecuencias** o **Percentages** y **Horizontal** o **Vertical**.

Diagramas de Cajas

Empleamos **Plot** seleccionando **Exploratory plots** y como opción **Box and Whisker plot** (existe la posibilidad de ir directamente, si está cargada la **Toolbar**, a través del botón que tiene el dibujo de una caja, el cual ocupa la posición central).

Si se quiere cambiar los parámetros se deberá pulsar dos veces sobre el gráfico y una vez ampliado pulsar en el interior primero con el habitual botón izquierdo del ratón y después con el derecho para finalmente seleccionar **Pane Options**.

Las **Pane Options** relevantes que se pueden marcar son: **Median Notch** (intervalo de confianza aproximado sobre la mediana), **Outliers Symbols** (dibuja los valores atípicos) y **Mean Marker** (señala la media).

Medidas de posición, dispersión y otras

Usamos **Describe** y seleccionamos **Categorical data** y posteriormente **One-variable analysis**. En el campo **Data** escribimos el nombre de la variable y pulsamos **OK**.

Aparecerá la ventana **One-variable analysis** con una barra de botones en la que si pulsamos el segundo por la izquierda (**tabular options**) tenemos:

Summary Statistics Colección de medidas descriptivas:

1. Media (Average)
2. Mediana (Median)
3. Moda (Mode)
4. Cuartiles (Quartiles): primero, segundo y tercero
5. Rango intercuartílico
6. Cuasivarianza (Variance)
7. Cuasidesviación típica (Std. deviation)
8. Coeficiente de Variación de Pearson (Coefficient of Variation)
9. Recorrido o rango muestral (Range)
10. Coeficiente de asimetría (Skewness)

11. Curtosis o apuntamiento

Percentiles Percentil que necesitemos (1%, 5%,..., 99%).

Frequency Tabulation Tabla de frecuencias con el número de clases deseado.

Confidence Intervals Intervalos de confianza con el nivel de confianza especificado para la media y desviación típica.

Si pulsamos el tercer botón (**graphical options**) podemos visualizar:

Box and Whisker Plot Otra posibilidad para obtener el diagrama de cajas.

Frequency Histogram Otra posibilidad para obtener el histograma.

Quantile Plot Dibuja los percentiles acumulados por cada valor.

DESCRIPCION CONJUNTA DE DOS VARIABLES

Tablas de frecuencias conjuntas

Las encontramos en el menú **Describe**, seleccionando **Categorical Data** y posteriormente **Crosstabulation**. En el campo **Row Variable** introducimos la variable que figurará en el eje horizontal, y en el campo **Column Variable** aquella que lo hará en el eje vertical, y finalmente pulsamos **OK**.

En la opción **Crosstabulation**, si pulsamos el segundo botón por la izquierda, se nos abrirá a su vez la ventana **Tabular Options**, en la que marcaremos el recuadro correspondiente a **Frequency Table**.

Histogramas tridimensionales

Vamos a la ventana **Crosstabulation**, como hemos indicado en el apartado anterior, y pulsamos el tercer botón por la izquierda, y dentro de **Graphical Options** escogemos **Skychart**.

Situando el ratón sobre el gráfico y pulsando su botón derecho, aparece la opción **Pane Options**, donde podemos escoger entre frecuencias y porcentajes.

Diagramas de dispersión

Los encontraremos en el menú **Plot**, submenú **Scatterplots**, seleccionando **X-Y plot**. Para continuar debemos rellenar los campos correspondientes a las variables dependiente e independiente.

También podremos acceder a él desde la barra de utilidades, pulsando el noveno botón por la izquierda, que se identifica con la etiqueta **Scatterplots**.

Regresión lineal simple

Nos situamos dentro del menú **Relate** y seleccionamos **Simple Regression**. Sólo nos interesarán algunas de las opciones que nos ofrece este submenú.

Rellenaremos los campos correspondientes a las variables dependiente e independiente.

Una vez abierta esta ventana, podemos ajustar nuestros datos a un modelo de relación lineal. Para ello, debemos pulsar el botón derecho del ratón y marcar **Analysis Options**.

De todos los resultados que aparecen en la pantalla, nos interesan: *Slope* (pendiente de la recta), *Intercept* (ordenada en el origen), *Correlation Coefficient* (coeficiente de correlación), *R-squared* (nos mide la bondad del ajuste).

Otras aplicaciones que nos pueden interesar son:

1. *Mostrar la recta ajustada y la nube de puntos*. Para ello, partimos de la ventana de **Simple Regression**, seleccionamos el tercer botón por la izquierda y aparecerá **Graphical Options**, donde marcamos **Plot of Fitted Model**.

2. *Hacer predicciones de valores de la variable dependiente*. Los pasos a realizar son los siguientes: nos situamos en la ventana de **Simple Regression**, y pulsando el segundo botón por la izquierda, seleccionamos dentro de **Tabular Options**, la opción **Forecast**. Sobre la pantalla que aparece, pulsamos el botón derecho del ratón y elegimos **Pane Options**, lo que nos permite obtener los valores previstos de la variable respuesta para valores determinados de la variable independiente.

Covarianza

Entramos en el menú **Describe**, seleccionamos **Numerical Data** y a continuación **Multiple-Variable Analysis**. En el campo **Data**, escribimos los nombres de las dos variables, y en la ventana obtenida, pulsamos el segundo botón de la izquierda marcando **Covariances** en **Tabular Options**. La información que se refleja en la pantalla, incluye también las varianzas de cada una de las variables.

DISTRIBUCIONES Y CALCULO DE PROBABILIDADES

Funciones de cuantía, densidad y distribución

Entramos en el menú **Plot**, y en él escogemos **Probability Distributions**. A continuación marcamos la función de probabilidad que queramos dibujar. En la ventana que aparece, pulsar el botón derecho del ratón, y marcar **Analysis Options**. Aquí podemos especificar la media y desviación típica con las que vayamos a trabajar. Si finalmente seleccionamos el ítem del menú **Graphical Options**, podemos escoger entre las siguientes opciones:

Density/Mass Function Dibuja la función de densidad o cuantía, según el caso.

CDF (Cumulative Density Function) Dibuja la función de distribución.

Cálculo de probabilidades con la función de distribución

Entramos en el menú **Plot**, y en él escogemos **Probability Distributions**, donde seleccionamos la función de distribución con la que vayamos a trabajar. Con el ítem de **Tabular Options** seleccionamos **Cumulative Distribution**. Una vez aquí, apretaremos el botón derecho del ratón, y en **Pane Options** encontramos **Cumulative Distribution Options**, donde pondremos los valores que nos pidan.

Cálculo de los valores críticos (p-valores)

Entramos en **Plot**, y seleccionamos **Probability Distributions**, donde hay que marcar la función de distribución con la que vayamos a trabajar. En la ventana obtenida, pulsar sobre el segundo botón, y escoger **Inverse CDF**. Sobre esta pantalla, apretamos el botón derecho del ratón, señalando **Pane Options**. Aparecerá la ventana **Inverse CDF Options**, y pondremos los valores de probabilidad de los que queramos calcular sus valores críticos.

Generación de números aleatorios

A partir de **Plot** vamos a **Probability Distributions**, y seleccionamos la función de distribución que nos interesa. En la ventana obtenida, pulsar sobre el segundo botón, y marcar **Random Numbers**. En la ventana que aparece, pulsamos el botón derecho del ratón y seleccionamos **Pane Options**, donde escogemos la cantidad de números aleatorios que queramos generar. Tras esta operación, pulsamos el ítem cuarto de la ventana, marcando **Random Numbers for Dist.1**. Así, esa cantidad de números aleatorios queda grabada para ser utilizada con posterioridad.

Nota: Para cambiar los límites de los ejes de las gráficas, hay que maximizar la gráfica, ir con el cursor a uno de los valores en el eje que queremos cambiar, y apretar el botón izquierdo del ratón; aparecen marcas en las esquinas del eje que queremos. Pulsamos el botón derecho del ratón, y apretamos sobre el botón izquierdo en **Axis Scaling Options**; completamos la ventana de diálogo con los números que queremos.

Si estamos en **StatGallery**, para llegar a **Axis Scaling Options** hay que pasar previamente por **Modify Item**.

INFERENCIA ESTADÍSTICA: INTERVALOS DE CONFIANZA Y CONTRASTES DE HIPÓTESIS

Intervalos de confianza y contrastes de hipótesis en una muestra

Vamos a trabajar dentro de la opción: **Describe/ Numeric Data/ One-Variable Analysis**. En la ventana que nos aparece, indicaremos el nombre de la variable.

Aparece la pantalla correspondiente a **Analysis Summary**, donde vemos reflejada el nombre de la variable, el número total de datos, y el rango de los mismos.

Si pulsamos el segundo de los cuatro iconos que nos aparecen en la parte superior de la pantalla (empezando por la izquierda), podemos activar las opciones **Confidence Intervals** y **Hypotesis Tests**.

- La opción de los *intervalos de confianza*, nos permite calcular los intervalos de confianza al 95% para la media y la desviación típica.

Si situamos el ratón sobre la ventana donde hemos obtenido los intervalos de confianza, y pulsamos el botón derecho, aparece la opción **Pane Options**, donde podremos cambiar el nivel de significación.

- La opción del *contraste de hipótesis*, da el resultado del contraste basado en la t-Student (donde la hipótesis nula por defecto es que la media de la variable es 0, y la hipótesis alternativa es que la media es distinta de 0). Además nos da el *p-valor*, y nos dice si se acepta o rechaza la hipótesis nula.

Si nos situamos sobre la ventana y pulsamos el botón derecho del ratón, aparece la opción **Pane Options**, donde podremos cambiar el valor que queremos contrastar de la media, el nivel de significación, y además nos permite tres tipos de hipótesis alternativa:

1. *Not Equal* La media es distinta del valor dado en la hipótesis nula.
2. *Less Than* La media es menor que el valor dado como hipótesis nula.
3. *Greater Than* La media es mayor que el valor dado en la hipótesis nula.

El tercero de los iconos (empezando por la izquierda) del margen superior izquierdo de la pantalla, nos permitirá realizar las siguientes representaciones gráficas:

Scatterplot Nube de puntos.

Box-and-Whisker Plot Diagrama de caja.

Histograma de frecuencias

Quantile Plot Dibuja los percentiles acumulados para cada valor.

Normal Probability Plot

Density Trace

Symmetry Plot

Intervalos de confianza y contrastes de hipótesis en dos muestras

En este caso vamos a trabajar dentro de la opción: **Compare/ Two Samples/ Two Sample Comparison**. Nos aparecerán los siguientes campos:

Sample 1 Pide el nombre de la variable correspondiente a la primera muestra.

Sample 2 Nombre de la variable correspondiente a la segunda muestra.

Select Permite seleccionar un subconjunto de las observaciones. Por ejemplo, si quisiéramos comparar los 50 primeros valores de cada variable pondríamos *first(50)* en este campo.

Sort Cuando está activado, ordena las variables alfabéticamente.

El segundo botón por la izquierda **Tabular Options**, nos permite seleccionar entre las siguientes seis opciones:

Analysis Summary

Summary Statistics Proporciona los principales estadísticos muestrales para las dos muestras por separado. Si pulsamos sobre la ventana el botón izquierdo del ratón, y escogemos la opción **Pane Options**, podemos activar la obtención del resto de estadísticos muestrales.

Comparison of Means Construye intervalos de confianza al 95% para la media de ambas variables, y para la diferencia entre las medias, primero asumiendo varianzas iguales, y luego suponiéndolas distintas. También nos realiza un contraste t-Student para la igualdad de medias frente a varias hipótesis alternativas. Para cambiar el nivel de significación recurriremos a la opción **Pane Options**.

Comparison of Standard Deviations Muestra la varianza y desviación típica de cada una de las variables, y construye sus intervalos de confianza al 95%. Calcula la proporción existente entre las dos varianzas, y el intervalo de confianza para este cociente también al 95%. Además realiza un contraste F para comparar las varianzas de las dos muestras. La hipótesis nula es la igualdad de las varianzas, y se contemplan las tres hipótesis alternativas posibles. El nivel de significatividad se puede cambiar desde la opción **Pane Options**.

Comparison of Medians Muestra las medianas de las dos distribuciones por separado, y realiza un contraste sobre la igualdad de las medianas.

Kolmogorov-Smirnov Test Realiza el contraste de Kolmogorov-Smirnov cuya hipótesis nula es que podemos suponer que las dos muestras proceden de la misma distribución.

2 PRACTICA 2

Introducción a los métodos estadísticos a través de demostraciones en páginas Web

La idea de esta práctica es recordar los conceptos que se han dado en el tema introductorio de teoría mediante unas demostraciones gráficas y numéricas de todo un conjunto de técnicas básicas en estadística.

Estas páginas se encuentran en la dirección Web <http://members.aol.com/johnp71/javastat.html/> titulada "Web Pages that Perform Statistical Calculations". En ellas hay una sección en la que aparecen todos los procedimientos básicos en estadística u otros de mayor nivel para esta práctica introductoria.

La idea es que el alumno, guiado por el profesor, entre en algunas direcciones y ejecute de forma interactiva los métodos correspondientes.

3 PRACTICA 3

Análisis de la Varianza con Statgraphics

COMANDOS NECESARIOS PARA ESTA PRACTICA

INTRODUCCION DE DATOS

En todos los procedimientos de ANOVA introduciremos los datos de la siguiente forma. En un fichero crearemos las siguientes variables: una primera variable que será la variable objeto de estudio y en la que introduciremos todos los datos de todos los niveles de los factores; después, tantas variables como factores tengamos y en ellas indicaremos a qué nivel de cada factor pertenece cada dato de la variable en estudio. En el caso de medidas repetidas o aleatorización en bloque añadiremos el factor sujeto o bloque indicando a qué sujeto o bloque pertenece cada observación.

ANOVA DE 1 VIA

El procedimiento lo encontramos en **compare**, en el submenú **Analysis of Variance**. Se llama **One-Way Anova** y seleccionamos los datos que queramos comparar,

Dependent variable Nombre de la variable continua respuesta.

Factor Nombre de la variable factor.

Tabular Options

Summary statistics Muestra los estadísticos descriptivos más importantes.

ANOVA table Calcula y muestra la tabla estándar del ANOVA de una vía.

Table of means Proporciona la media muestral y la desviación típica para cada nivel del factor. También nos da un intervalo de confianza para la media de cada nivel del factor.

Multiple Range Test Test de comparaciones múltiples con el método y nivel introducido.

Variance Check tabulation Estadísticos asociados a los contrastes de Cochran, Bartlett y Hartley para comprobar la homogeneidad de varianzas.

Graphical Options

Scatterplot Dibuja los valores de la variable respuesta para cada uno de los niveles del factor.

Means Plot Muestra la tabla de medias en formato gráfico.

Box-and-Whisker Plot Dibuja un diagrama de caja para los datos de cada nivel del factor.

Residuals vs Factor Levels Gráfico de residuos para cada nivel del factor. Es útil para ver si la varianza es constante entre niveles.

Residuals vs Predicted Gráfico de residuos frente valores predichos. Es útil para transformar la variable respuesta, cuando la varianza no es constante entre niveles.

Residuals vs Row Number Gráfico de residuos frente al orden de las observaciones. Es útil para identificar correlación serial.

ANOVA MULTIFACTORIAL

Aunque en teoría estudiamos sólo el caso de dos factores, con StatGraphics podemos analizar diseños de hasta 10 factores. El procedimiento lo encontramos en **compare**, en el submenú **Analysis of Variance**. Se llama **Multifactor Anova** y seleccionamos los datos que queramos comparar,

Dependent variable Nombre de la variable continua respuesta.

Factors Nombre de los factores.

Covariates Nombre de hasta 3 covariables.

Tabular Options

Analysis Summary Muestra los nombres de la variable dependiente, los factores, covariables y el número de casos de estudio. Con **Options** completamos **Maximum Order Interaction** y **Exclude**.

ANOVA table Calcula y muestra la tabla estándar del ANOVA con las siguientes opciones:

- *Suma de cuadrados*. Type I, para diseños balanceados y Type III para ambos, balanceados y no.
- *Factor*. Factor para el que queremos seleccionar el término del error.
- *Error Term*. Tipo de media cuadrática para el cálculo de la F. Por defecto utiliza el residual.
- *Selections*. Nos permite ver la selección realizada.

Table of means Proporciona la tabla de medias mínimo cuadráticas y los intervalos de confianza asociados.

Multiple Range Test Test de comparaciones múltiples con el método y nivel introducido.

Graphical Options

Scatterplot Dibuja los valores de la variable respuesta para cada uno de los niveles del factor seleccionado.

Means Plot Muestra la tabla de medias para el factor seleccionado en formato gráfico.

Interaction Plot Crea un gráfico que muestra las interacciones entre dos factores seleccionados.

Residuals vs Factor Levels Gráfico de residuos para cada nivel del factor seleccionado. Es útil para ver si la varianza es constante entre niveles.

Residuals vs Predicted Gráfico de residuos frente valores predichos. Es útil para transformar la variable respuesta, cuando la varianza no es constante entre niveles de un factor seleccionado.

Residuals vs Row Number Gráfico de residuos frente al orden de las observaciones. Es útil para identificar correlación serial.

ANOVA DE MEDIDAS REPETIDAS Y DISEÑO EN BLOQUES

Podemos realizar un diseño de medidas repetidas utilizando el procedimiento del anova multifactorial (podemos considerar medidas repetidas en sólo uno, varios o todos los factores). Los campos de entrada son:

Factors El factor sujeto es un factor más. El orden será el siguiente: factores sin medidas repetidas, factor sujeto, factores con medidas repetidas.

Maximum order interaction Este número será como máximo el número de factores diferentes al factor sujeto. Para el factor sujeto, dejamos el término del error vacío. Para los factores sin medidas repetidas, el término del error será el factor sujeto. Para el resto, este término será el residuo.

El resto es como en el que caso de **Anova Multifactorial**.

EJERCICIOS

1. Cuando un lenguaje de alto nivel es compilado, el tiempo de ejecución depende del compilador. Un ingeniero del Software desea comparar tres compiladores que han sido desarrollados para un nuevo lenguaje de alto nivel. Quince programas son seleccionados de estos que se consideran apropiados para el nuevo lenguaje y cada uno es codificado en el nuevo lenguaje por el mismo programador, cinco son aleatoriamente seleccionados para cada compilador. Los tiempos de CPU fueron:

Compilador A: 8.4, 17.4, 9.8, 8, 15.3
Compilador B: 9.2, 5.8, 14.3, 13.5, 7
Compilador C: 4.7, 9.6, 10, 18.3, 6.7

Se pide:

- a) Tabla ANOVA.
- b) Determinar a nivel 0.05 y a nivel 0.01 si la hipótesis nula de igualdad de tiempos medios puede ser rechazada.
- c) Comprobar la homogeneidad de varianzas.

2. Un ingeniero experto en tratamiento de imágenes desea determinar si hay diferencia entre el efecto de dos algoritmos de reconocimiento A y B. Antes de la aplicación, la imagen debe ser filtrada para eliminar el ruido, y para ello puede elegir entre dos filtros F y G. Además las imágenes son creadas ópticamente en un nivel alto o bajo de luz. Se analizan 24 imágenes con los siguientes ratios de error:

	F		G	
	bajo	alto	bajo	alto
A	67	78	70	64
	63	74	70	62
	65	70	67	60
B	40	50	58	47
	48	54	60	50
	44	54	59	53

- Determinar qué factores e interacciones son significativos a nivel 0.01.
- Tabla ANOVA.
- Intervalos de confianza para las medias en los diferentes algoritmos.
- Compara el algoritmo A con el filtro F y el algoritmo B con el filtro G.

3. Doce personas son distribuidas aleatoriamente en cuatro grupos de tres personas cada uno de ellos . A cada grupo, se le asigna aleatoriamente un tiempo distinto de entrenamiento antes de verificar cierta tarea programada en ordenador. Los resultados en dicha tarea, con los correspondientes tiempos de entrenamiento, vienen dados en la siguiente tabla:

	0.5 horas	1 hora	1.5 horas	2 horas
1	4	3	8	
3	6	5	10	
5	2	7	6	

Con $\alpha = 0.05$, ¿es compatible con los resultados experimentales la hipótesis nula, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$?

4. El método de fluorescencia de Rayos X es una herramienta analítica importante para determinar la concentración del material de propulsores sólidos para misiles. Se postuló que el proceso de mezcla de propulsor y el momento de análisis tiene una influencia sobre la homogeneidad del material y por lo tanto en la precisión de la intensidad de rayos X. Se llevó a cabo un experimento utilizando tres factores: A=condiciones del mezclado (cuatro niveles), B=momento del análisis (dos niveles) y C=método de colocación del propulsor en los soportes de la mezcla (caliente y temperatura ambiente). Se registraron los siguientes datos los cuales representan el análisis en porcentaje de peso de perclorato de amonio en un propulsor particular. Realiza un análisis de varianza con $\alpha = 0.01$ para probar la significatividad de los efectos principal y de interacción.

	C1		C2	
	B1	B2	B1	B2
1	38.62	38.45	39.82	39.82
	37.20	38.64	39.15	40.26
	38.02	38.75	39.78	39.72
2	37.67	37.81	39.53	39.56
	37.57	37.75	39.76	39.25
	37.85	37.91	39.90	39.04
3	37.51	37.21	39.34	39.74
	37.74	37.42	39.60	39.49
	37.58	37.79	39.62	39.45
4	37.52	37.60	40.09	39.36
	37.15	37.55	39.63	39.38
	37.51	37.91	39.67	39.00

5. Un taller desea adquirir un nuevo modelo de torno. Los fabricantes ofertan tres tipos de máquinas. Para probarlas, el jefe de taller dispone que tres operarios trabajen con los tres tipos de máquinas durante cinco jornadas, asignando el orden de trabajo en cada máquina al azar. Después de la experiencia se obtuvieron los siguientes resultados (número de piezas fabricadas por jornada):

Maq.	Oper. 1					Oper. 2					Oper. 3				
1	52	54	49	51	52	55	54	57	50	51	47	48	47	46	44
2	59	61	62	65	57	60	60	61	63	58	57	56	59	54	50
3	44	43	48	47	50	45	49	48	51	51	39	41	47	36	52

Examinar si existen diferencias entre máquinas, entre operarios y la significación de la interacción máquina-operario. ¿Qué máquina sería la más productiva?

6. A menudo, un procedimiento numérico puede ser obtenido usando diferentes algoritmos. Consideremos la comparación de tres algoritmos, cada uno codificado en un lenguaje de alto nivel. Como el tiempo de ejecución se ve muy afectado por la elección del hardware, cada algoritmo es ejecutado en uno de cuatro ordenadores diferentes. Los datos son:

	Algoritmo 1	Algoritmo 2	Algoritmo 3
Ord. 1	6.42	6.85	5.60
Ord. 2	10.51	11.45	9.50
Ord. 3	4.00	4.12	3.94
Ord. 4	6.20	6.72	5.66

- ¿Existen diferencias significativas entre los tres algoritmos? ¿Y entre los ordenadores?
- ¿Qué algoritmo proporciona una media mayor? ¿Podemos decir que es el mejor?

4 PRACTICA 4

Análisis de Regresión con Statgraphics

COMANDOS NECESARIOS PARA ESTA PRACTICA

REGRESION SIMPLE

El procedimiento lo encontramos en **Relate**, en el submenú **Simple Regression**. Seleccionamos los campos siguientes:

Dependent variable Nombre de la variable continua respuesta.

Independent variable Nombre de la variable explicativa.

Model El procedimiento de regresión simple ajusta un modelo en el que relaciona la variable dependiente con la variable respuesta mediante minimización de la suma de cuadrados de los residuos de la curva ajustada. Este procedimiento permite ajustar los siguientes modelos:

- *Linear* $y = bx + a$
- *Multiplicative* $y = ax^b$
- *Exponential* $y = \exp(bx + a)$
- *Reciprocal* $\frac{1}{y} = a + bx$

Tabular Options

Analysis Summary Muestra los estadísticos descriptivos más importantes.

Forecast Valores estimados para una nueva predicción y/o la media para un valor de x dado. También da intervalos de confianza para ambos valores.

Unusual Residuals Observaciones con residuos atípicos.

Influential Points Detecta las observaciones influyentes a la hora de construir el modelo.

Graphical Options

Plot fitted model Muestra la curva ajustada y la nube de puntos. Permite introducir un valor de la variable independiente (en la parte inferior del gráfico) seguido de una coma y nos dibuja el valor de la variable respuesta. Inversamente, dada una variable respuesta, precedida de una coma, nos da el valor de la variable independiente asociado.

Plot observed versus predicted Dibujo de los residuos en función de los valores predichos.

Plot residuals versus x Dibujo de los residuos en función de la variable explicativa.

Plot residuals versus row number Dibujo de los residuos en función del orden de las observaciones.

REGRESION MULTIPLE

El procedimiento lo encontramos en **Relate**, en el submenú **Multiple Regression**. Seleccionamos los siguientes campos:

Dependent variable Nombre de la variable continua respuesta.

Independent variable Nombre de las variables explicativas.

Una vez especificados los campos, comienza el análisis del modelo. En **Analysis Options** podemos elegir las técnicas backward o forward.

Tabular Options

Analysis Summary Muestra los valores ajustados del modelo y los estadísticos para contrastar su significatividad. Además, muestra la tabla ANOVA con el coeficiente de determinación.

Conditional Sum of Squares Descompone la suma de cuadrados debida al modelo en sumas de cuadrados, cada uno de ellos debidos a cada una de las variables explicativas.

Confidence Intervals Intervalos de confianza para los parámetros del modelo.

Correlation Matrix Matriz de correlaciones entre los estimadores de los parámetros del modelo.

Reports Para cada observación, calcula los valores observados, predichos, residuos e intervalos de confianza para la media y predicción.

Unusual Residuals Observaciones con residuos atípicos.

Influential Points Detecta las observaciones influyentes a la hora de construir el modelo.

Graphical Options

Plot observed versus predicted Dibujo de los residuos en función de los valores predichos.

Plot residuals versus x Dibujo de los residuos en función de la variable explicativa.

Plot residuals versus row number Dibujo de los residuos en función del orden de las observaciones.

Interval Plots Dibuja distintos intervalos de confianza seleccionados a través de **pane options**.

EJERCICIOS

1. La siguiente tabla proporciona la velocidad máxima (V), la potencia en CV (P), la cilindrada en cc (C) y el número de cilindros de una muestra de 24 motos distintas que pueden adquirirse en España:

V	P	C	N
160	60	247	2
156	26	231	4
193	80	906	6
191	67	980	2
200	70	980	2
190	70	980	2
194	70	980	2
170	36	360	1
132	27	325	1
111	17	249	1
177	50	493	2
173	63	626	4
115	16	178	1
203	83	1085	4
111	17	249	1
170	27	498	1
156	44	395	2
115	17	246	2
137	27	443	2
195	62	553	4
220	86	1116	3
120	17	124	2
105	23	250	1
85	7	124	1

- a) Ajustar un modelo lineal a la relación entre la velocidad y las otras variables. ¿Son todos los coeficientes significativamente distintos de cero a nivel 0.05?
- b) Hallar un intervalo de confianza para el coeficiente de la potencia al 99%.
- c) Analizando los residuos, ¿es apropiado el modelo?
- d) Dar un intervalo de confianza al 95 % para la predicción en (60,200,2).

2. Los datos siguientes relacionan la temperatura de ebullición del agua en grados Fahrenheit, con la presión barométrica, y fueron tomados por el físico escocés Forbes en 1857 en los Alpes y en Escocia. Se pide:

- a) Construir un modelo para prever la temperatura (y) en función de la presión (x).
- b) ¿Es la pendiente significativa a nivel 0.01?
- c) Calcular el coeficiente de correlación.
- d) Dar un intervalo de confianza al 90 % para la temperatura media a una presión de 25.

Presión: 20.79, 22.40, 23.15, 23.89, 24.02, 25.14, 28.49, 29.04, 29.88, 30.06
 Temperatura: 194.5, 197.9, 199.4, 200.9, 201.4, 203.6, 209.5, 210.7, 211.9, 212.2

3. Se realiza un experimento para determinar la duración de vida de ciertos circuitos electrónicos (y) en función de dos variables de fabricación (x_1 , x_2) con los resultados siguientes:

y	x_1	x_2
11	-10	0
8	0	-5
73	10	5
21	-10	0
46	0	5
30	10	-5

Se pide:

- a) Calcular la ecuación de regresión y determinar el valor de R^2 .
- b) Construir un intervalo de confianza al 95% para la predicción en el punto (0,0).
- c) Comentar la validez de las hipótesis previas utilizando los residuos.

4. Una agencia desea estimar los gastos de alimentación de una familia con base al ingreso y su tamaño. Los datos que se encuentran en la siguiente tabla representan los gastos de alimentación por mes (Y) en miles de dólares, el ingreso mensual (X_1), y el tamaño de la familia (X_2) para 15 familias.

y	x_1	x_2
0.43	2.1	3
0.31	1.1	4
0.32	0.9	5
0.46	1.6	4
1.25	6.2	4
0.44	2.3	3
0.52	1.8	6
0.29	1	5
1.29	8.9	3
0.35	2.4	2
0.35	1.2	4
0.78	4.7	3
0.43	3.5	2
0.47	2.9	3
0.38	1.4	4

- a) Ajustense todos los modelos lineales que abarcan a X_1 , y/o X_2 , e interpreten los coeficientes de regresión estimados.
- b) Probar la hipótesis nula $H_0 : b_1 = b_2 = 0$.
- c) Calcular e interpretar el coeficiente de determinación.
- d) Con base en los resultados anteriores, decídase cuál es la mejor ecuación para predecir el gasto de alimentación y empléese para estimar el gasto promedio mensual en alimentación para una familia de cuatro personas con un ingreso mensual de 2500. Determinese un intervalo de confianza al 98% para esta cantidad.

Análisis discriminante con Statgraphics

COMANDOS NECESARIOS PARA ESTA PRACTICA

INTRODUCCION DE DATOS

En el procedimiento de Análisis Discriminante introduciremos los datos de la siguiente forma. En un fichero crearemos las siguientes variables: una primera variable que indicará a qué población o clase pertenece cada individuo. Después, tantas variables como hayamos medido a cada individuo sin tener en cuenta la población.

PROCEDIMIENTOS

Los comandos los encontramos dentro del submenú **Multivariate Methods** y el procedimiento se llama **Discriminant Analysis**. Los siguientes campos tienen que ser rellenados:

Classification Factor Nombre de la variable que define las clases o poblaciones.

Data Variables Resto de variables.

Además, tenemos las siguientes opciones en **tabular options**:

Analysis Summary Proporciona los valores propios de la matriz $W^{-1}B$ distintos de cero, el porcentaje de discriminación que representan y si la transformación asociada es o no significativa.

Display Standardized Coefficients Proporciona los coeficientes de la transformación a variedades canónicas, si las variables estuvieran tipificadas. Es útil para ver qué variables influyen más en la variedad canónica y por tanto en la discriminación.

Display Unstandardized Coefficients Proporciona los coeficientes de la transformación a variedades canónicas y que utilizaremos para clasificar una nueva observación.

Display Group Centroids Muestra los centroides de cada grupo o población.

Display Group Statistics Muestra las medias y desviaciones típicas de las variables originales en cada grupo o población.

Display Group Correlations Muestra la matriz W y W dividida por las correspondientes desviaciones típicas.

Classify Observations Los datos originales son reclasificados utilizando las variedades canónicas y nos da una medida de fiabilidad indicando porcentajes de clasificaciones correctas.

Plot Discriminant Functions Dibuja los datos transformados (aunque sólo lo hace para el caso de dos variedades canónicas.).

1. Existen tres tipos de fármacos ansiolíticos (A, B, C), la principal diferencia está en que cuando los aplicamos producen efectos secundarios diferentes en la conducción. Tenemos las siguientes observaciones:

Individuos con far. A:

x_1 : 0.548, 0.619, 0.641, 0.628, 0.846, 0.517, 0.876, 0.602

x_2 : 177.8, 184.4, 247.2, 163.4, 173.6, 167.2, 174, 158.6

Individuos con far. B:

x_1 : 0.519, 0.776, 0.678, 0.595, 0.858, 0.493, 0.741, 0.719

x_2 : 203, 164.8, 215.8, 153.6, 171.6, 166, 170.2, 157.2

Individuos con far. C:

x_1 : 0.637, 0.818, 0.701, 0.687, 0.855, 0.618, 0.849, 0.731

x_2 : 194.8, 175.2, 250.8, 152.2, 189.2, 181, 189, 184.6

donde x_1 =tiempo de reacción al ponerse un semáforo rojo, x_2 =apreciación de distancias (distancia mínima entre dos postes para poderlos atravesar).

a) Escribe los centroides de cada grupo (sujetos con cada uno de los fármacos) y la matriz W .

b) Si un individuo tarda 0.7 seg en reaccionar y tiene una apreciación de 180 cm., ¿qué fármaco ha tomado? ¿Qué fiabilidad darías a este resultado?

c) ¿Qué variable influye más en la clasificación?

2. Una empresa de microprocesadores dispone de dos máquinas especiales para ejecutar una tarea en la producción de una cierta marca de procesadores. A estos procesadores se les mide un par de características X_1 y X_2 porque se quiere poder distinguir entre máquinas en función de estas características. Los datos fueron los siguientes:

Maq.1		Maq.2	
X_1	X_2	X_1	X_2
191	131	186	107
185	134	211	122
200	137	201	114
173	127	242	131
171	118	184	108
160	118	211	118
188	134	217	122
186	129	223	127
174	131	208	125
163	115	199	124
190	143	211	129
174	131	218	126
201	130	203	122
190	133	192	116
182	130	195	123
184	131	211	122
177	127	187	123
178	126	192	109

Se pide:

- Calcula los centroides y la matriz W .
- Clasificar la nueva observación (190,125).
- Determinar la variable que más influye en la clasificación.

3. Llevar a cabo un análisis discriminante de los tres grupos siguientes (A , B , C) en los cuales se observan tres variables X_1 , X_2 y X_3 . Los datos fueron:

A			B			C		
X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
6	7	4	11	13	30	18	16	25
7	5	1	15	16	42	24	22	30
9	10	11	22	20	50	20	21	32
8	8	6	17	16	45	19	20	27
8	9	6	12	11	38	22	25	34
10	9	8	13	14	35	17	16	32

Se pide:

- Obtener los valores propios de la matriz $W^{-1}B$, discute el porcentaje de variabilidad que representan y determinar si la transformación asociada es o no significativa.
- Transformación a variedades canónicas.
- Dibujar los datos transformados.

6 PRACTICA 6

Análisis cluster con Statgraphics

COMANDOS NECESARIOS PARA ESTA PRACTICA

INTRODUCCION DE DATOS

En el procedimiento de Análisis Cluster introduciremos, como normalmente lo hacemos, las variables por columnas y las observaciones por filas. Si necesitamos definir matrices de distancias, las operaciones relacionadas con matrices las podemos encontrar en **Matrix Utilities**.

PROCEDIMIENTOS

Los comandos los encontramos dentro del submenú **Multivariate Methods** y el procedimiento se llama **Cluster Analysis**. Los siguientes campos tienen que ser rellenados:

Data Nombre de la matriz donde están almacenadas las distancias o bien conjunto de variables que queremos considerar.

Labels Nombre de la variable cualitativa donde hemos almacenado los nombres (si los hay) de cada observación.

Method Métodos para la construcción de clusters (cada método viene explicado en teoría):

- *Average*
- *Centroid*
- *Furthest*
- *Nearest*
- *Median*

Distance Si partimos de una matriz de distancias, seleccionaremos **Matrix** y en otro caso si queremos utilizar las distancias Euclídeas, seleccionaremos **Euclidean**.

Clusters Número de clusters que queremos formar.

Standardize Estandariza las variables

En lo que respecta a los gráficos, también disponemos de algunas opciones como son:

X-axis Variable que queramos que aparezca en el eje de las X para dibujar los clusters.

Y-axis Variable que queramos que aparezca en el eje de las Y para dibujar los clusters.

Z-axis Variable que queramos que aparezca en el eje de las Z para dibujar los clusters.

Plot Cluster Dibuja los clusters con las variables en cada eje (según han sido especificadas). Este comando sólo se puede ejecutar si accedemos al análisis cluster mediante variables y no matriz de distancias.

EJERCICIOS

1. Los siguientes datos corresponden a 10 cráneos de gorila que se encuentran en el museo británico. Se cree que estos cráneos pertenecen a dos especies diferentes. En cada cráneo tres medidas fueron tomadas y los resultados fueron:

- x_1 , la longitud basal excluyendo la premaxila,
- x_2 , la longitud occipito nasal,
- x_3 , la longitud máximo nasal.

x_1	x_2	x_3
2.088	2.090	1.580
2.088	2.1	1.552
2.090	2.090	1.613
2.045	2.054	1.673
2.056	2.068	1.703
2.097	2.093	1.563
2.117	2.125	1.563
2.140	2.146	1.651
2.070	2.073	1.703
2.051	2.094	1.693

Se pide:

- a) Clasificar en dos grupos utilizando el método del centroide. Describe, ayudándote de algún gráfico, los dos grupos. Guarda la clasificación en una variable.
- b) Utilizando la clasificación obtenida en a) a qué especie pertenecería un cráneo con medidas (2.1,2.0,1.6) ? Qué fiabilidad darías a este resultado? Qué variable influye más en la clasificación?

2. Los siguientes datos corresponden a empresas españolas :

Ventas	Capital	Plantilla
3.14	0.967	810
4.735	2.437	944
4.5351	4.509	1005
2.5027	1.235	473
2.984	0.7713	2317
4.583	0.621	3087
3.097	2.002	3030
3.072	2.678	344
1.994	0.41	1046
2.007	8.726	5074

Se pide:

- a) Clasificar en tres grupos utilizando el método de la distancia al vecino más lejano.
- b) Describe, ayudándote de algún gráfico, los tres grupos.
- c) Clasificar en dos grupos utilizando un método no jerárquico y sabiendo que las dos primeras observaciones están en grupos diferentes.

3. Los siguientes datos corresponden a 20 muestras de suelos. Clasifícalos en cuatro grupos utilizando los métodos del vecino más cercano, del más lejano y de la media. Dibuja los cuatro grupos obtenidos respecto a las variables contenido de arena y materia orgánica.

% Arena	% Sedimento	% Arcilla	% M. Orgánica	pH
77.3	13	9.7	1.5	6.4
82.5	10.0	7.5	1.5	6.5
66.9	20.6	12.5	2.3	7.0
47.2	33.8	19.0	2.8	5.8
65.3	20.5	14.2	1.9	6.9
83.3	10.0	6.7	2.2	7.0
81.6	12.7	5.7	2.9	6.7
47.8	36.5	15.7	2.3	7.2
48.6	37.1	14.3	2.1	7.2
61.6	25.5	12.9	1.9	7.3
58.6	26.5	14.9	2.4	6.7
69.3	22.3	8.4	4	7
61.8	30.8	7.4	2.7	6.4
67.7	25.3	7	4.8	7.3
57.2	31.2	11.6	2.4	6.5
67.2	22.7	10.1	3.3	6.2
59.2	31.2	9.6	2.4	6
80.2	13.2	6.6	2	5.8
82.2	11.1	6.7	2.2	7.2
69.7	20.7	9.6	3.1	5.9

4. Un botánico afirma que existen dos especies claramente diferenciadas de nogales y que se pueden distinguir por la forma de sus hojas. Para determinar estas dos especies, mide la anchura y longitud en una muestra de 16 hojas de nogal. Los resultados fueron:

Hoja	Anchura	Longitud	Hoja	Anchura	Longitud
1	2.1	4.1	9	5.9	7.2
2	2.4	6.0	10	6.6	13.1
3	3.6	5.5	11	7.4	11.3
4	3.7	8.2	12	8.2	15.6
5	4.3	7.5	13	8.8	13.4
6	5.1	12.6	14	9.0	19.0
7	5.5	8.1	15	9.1	15.8
8	5.8	10.8	16	9.8	14.6

- a) ¿A qué especie pertenecería cada hoja? Utiliza en tu respuesta un método jerárquico.
- b) Representa gráficamente los resultados.

7 PRACTICA 7

Análisis de componentes principales con Statgraphics

COMANDOS NECESARIOS PARA ESTA PRACTICA

El método de componentes principales lo encontramos en la opción **Multivariate Methods** y **Principal Components**. El objetivo de este procedimiento es reducir lo máximo posible el número de variables y esto lo hace buscando combinaciones lineales de las variables originales del banco de datos.

Para su ejecución, rellenamos el campo **Data** con las variables de nuestro banco de datos y cuya dimensión queremos reducir. Para llevar a cabo el análisis, disponemos de tabular y graphical options.

Tabular Options

Analysis Summary Proporciona el valor de cada valor propio así como el porcentaje de varianza explicada por cada uno de ellos. Con **Pane Options** podemos modificar el criterio de obtención del número de componentes principales.

Component Weights Proporciona las ecuaciones de dichas componentes principales calculadas sobre los valores estandarizados de las variables originales.

Data Table Proporciona para cada observación sus correspondientes coordenadas en las componentes principales.

Graphical Options

Scree Plot Muestra el porcentaje de variación o el valor del valor propio para cada componente.

2D y 3D Scatterplot Muestra los valores de las componentes principales.

2D y 3D Component plot Muestra los pesos de cada variable en las correspondientes componentes.

2D y 3D Biplot Plot que proporciona la misma información que el Scatterplot y Component plots en un mismo gráfico. Las líneas representan las localizaciones de las variables en el espacio de las componentes.

EJERCICIOS

1. Realizar análisis de componentes principales con los siguientes bancos de datos procedentes del propio Statgraphics:

- Archivo de datos: CARDATA
- Archivo de datos: HOUSES
- Archivo de datos: MVDATA

2. Supongamos que una empresa necesita formar una medida de índice de precio al consumo (IPC) a partir de una colección de variables medidas en diferentes ciudades. Los datos vienen representados en la tabla adjunta. Realizar un análisis de componentes principales para formar este índice.

Ciudad	Pan	Tomates	Leche	Naranjas
1	24.5	41.6	73.9	80.1
2	26.5	53.3	67.5	74.6
3	29.7	59.6	61.4	104.0
4	22.8	51.2	65.3	118.4
5	26.7	51.2	62.7	105.9
6	25.3	45.6	63.3	99.3
7	22.8	46.8	52.4	110.9
8	23.3	41.8	62.5	117.9
9	24.1	52.4	51.5	109.7
10	29.3	61.7	80.2	133.2
11	22.3	42.4	67.8	108.6
12	26.1	43.2	65.4	100.9
13	26.9	38.4	56.2	82.7
14	20.3	53.9	53.8	111.8
15	24.6	50.7	51.9	106.0
16	30.8	62.6	66.0	107.3
17	24.5	61.7	66.7	98.0
18	26.2	49.3	60.2	117.1
19	26.5	46.2	60.8	115.1

3. Los siguientes datos corresponden a 20 muestras de suelos. Buscar las componentes principales asociadas a las 5 variables características de suelos.

% Arena	% Sedimento	% Arcilla	% M. Orgánica	pH
77.3	13	9.7	1.5	6.4
82.5	10.0	7.5	1.5	6.5
66.9	20.6	12.5	2.3	7.0
47.2	33.8	19.0	2.8	5.8
65.3	20.5	14.2	1.9	6.9
83.3	10.0	6.7	2.2	7.0
81.6	12.7	5.7	2.9	6.7
47.8	36.5	15.7	2.3	7.2
48.6	37.1	14.3	2.1	7.2
61.6	25.5	12.9	1.9	7.3
58.6	26.5	14.9	2.4	6.7
69.3	22.3	8.4	4	7
61.8	30.8	7.4	2.7	6.4
67.7	25.3	7	4.8	7.3
57.2	31.2	11.6	2.4	6.5
67.2	22.7	10.1	3.3	6.2
59.2	31.2	9.6	2.4	6
80.2	13.2	6.6	2	5.8
82.2	11.1	6.7	2.2	7.2
69.7	20.7	9.6	3.1	5.9

8 PRACTICA 8

Análisis factorial con Statgraphics

COMANDOS NECESARIOS PARA ESTA PRACTICA

El análisis factorial implementado en Statgraphics utiliza el método de componentes principales para extraer los factores. El procedimiento permite seleccionar el número de factores a extraer y puede utilizar diferentes métodos de rotación matricial.

Lo encontramos dentro de la opción **Multivariate Methods** en el submenú **Factor Analysis**. Tenemos los siguientes campos:

Data variables or filename Escribimos el nombre de las variables originales sobre las que se desea realizar un análisis factorial. Si se dispone de la matriz de correlación o covarianzas de las variables originales ésta se introduce en el campo Correlation or Covariance matrix.

Standardize Para realizar un análisis factorial hemos de estandarizar las variables para eliminar los efectos particulares de cada una de ellas. Para ello ponemos Yes en este campo.

Type of rotation Normalmente se utiliza la rotación Varimax aunque son perfectamente válidas las otras dos opciones de rotación.

Select Plot Factor Weights Opción que dibuja los pesos de cada variable en los factores extraídos tanto para la matriz factorial original como para la matriz una vez aplicado el procedimiento de rotación. Sobre el dibujo podemos identificar las coordenadas con el nombre de la variable original. Para ello seleccionamos Special del menú gráfico y posteriormente Identify.

EJERCICIOS

1. Realizar análisis factoriales con los siguientes bancos de datos procedentes del propio Statgraphics:

- Archivo de datos: CARDATA
- Archivo de datos: HOUSES
- Archivo de datos: MVDATA
- Archivo de datos: HOSPITAL
- Archivo de datos: PRESS88

2. Los siguientes datos corresponden a 20 muestras de suelos. Realiza un completo análisis factorial sobre las 5 variables para reducir la dimensión del problema en cuestión.

% Arena	% Sedimento	% Arcilla	% M. Orgánica	pH
77.3	13	9.7	1.5	6.4
82.5	10.0	7.5	1.5	6.5
66.9	20.6	12.5	2.3	7.0
47.2	33.8	19.0	2.8	5.8
65.3	20.5	14.2	1.9	6.9
83.3	10.0	6.7	2.2	7.0
81.6	12.7	5.7	2.9	6.7
47.8	36.5	15.7	2.3	7.2
48.6	37.1	14.3	2.1	7.2
61.6	25.5	12.9	1.9	7.3
58.6	26.5	14.9	2.4	6.7
69.3	22.3	8.4	4	7
61.8	30.8	7.4	2.7	6.4
67.7	25.3	7	4.8	7.3
57.2	31.2	11.6	2.4	6.5
67.2	22.7	10.1	3.3	6.2
59.2	31.2	9.6	2.4	6
80.2	13.2	6.6	2	5.8
82.2	11.1	6.7	2.2	7.2
69.7	20.7	9.6	3.1	5.9

9 PRACTICA 9

Series temporales con Statgraphics

COMANDOS NECESARIOS PARA ESTA PRACTICA

Todas las técnicas relacionadas con series temporales las encontramos dentro del submenú **Time Series Analysis**. Los procedimientos son:

- *Horizontal time sequence plot* dibuja los datos en función del tiempo (permite dibujar hasta doce series a la vez). Tenemos los siguientes campos:
 - *Y1-Y12*: escribimos el nombre de las variables donde guardamos los datos de las series.
 - *Time units*: usamos este campo para elegir las unidades de tiempo que queremos que el sistema use en el eje de las X.
- *Differencing* calcula la diferencia de la serie, puede ser simple o estacional. Tenemos los siguientes campos:
 - *Data*: escribimos el nombre de la variable donde guardamos los datos de la serie.
 - *Time units*: usamos este campo para especificar la longitud del ciclo estacional. Si especificamos un uno será una diferencia simple.
- *Box-Cox transformation* aplica la transformación de Box-Cox a los datos. Tenemos los siguientes campos:
 - *Data*: escribimos el nombre de la variable donde guardamos los datos de la serie.
 - *Transformation*: usamos este campo para elegir el tipo de transformación: normal o inversa.
 - *Lambda1*: introducimos el valor de lambda1.
 - *Lambda2*: introducimos el valor de lambda2.
 - *Geometric Mean*: introducimos la media geométrica (sólo hace falta para la transformación inversa).
- *Partial Autocorrelation function* calcula y dibuja la función de autocorrelación parcial. Tenemos los siguientes campos:
 - *Data*: escribimos el nombre de la variable donde guardamos los datos de la serie.
 - *Number of lags*: introducimos el retardo máximo para el cual el sistema calculará la función de autocorrelación parcial.

Otras opciones del sistema son:

- *Plot PACF*: dibuja la función de autocorrelación parcial como función del retardo.
- *Display Table*: muestra una tabla con los coeficientes estimados y los errores estándar.
- *Save PACF*: guarda los coeficientes estimados en un fichero.

- *Autocorrelation function* calcula y dibuja la función de autocorrelación. Tenemos los siguientes campos:

- *Data*: escribimos el nombre de la variable donde guardamos los datos de la serie.
- *Number of lags*: introducimos el retardo máximo para el cual el sistema calculará la función de autocorrelación.

Otras opciones del sistema son:

- *Plot ACF*: dibuja la función de autocorrelación como función del retardo.
 - *Display Table*: muestra una tabla con los coeficientes estimados y los errores estándar.
 - *Save ACF*: guarda los coeficientes estimados en un fichero.
- *Box-Jenkins ARIMA Modeling*: estimación y diagnóstico del modelo y predicción usando la metodología de Box and Jenkins. Tenemos los siguientes campos:
 - *Time series*: escribimos el nombre de la variable donde guardamos los datos de la serie.
 - *Order of Nonseasonal AR Factor*: introducimos el orden del término AR no estacional.
 - *Order of Nonseasonal Difference*: introducimos el orden de la diferencia no estacional que queremos que el sistema aplique para estimar el modelo.
 - *Order of Nonseasonal MA Factor*: introducimos el orden del término MA no estacional.
 - *Constant*: usamos este campo para elegir si queremos que el sistema incluya una constante en el modelo.
 - *Number of Forecasts*: usamos este campo para introducir el número de puntos que queremos que el sistema incluya en la predicción del modelo ajustado.
 - *Confidence Level*: nivel de confianza en el cálculo de los límites de predicción.
 - *Maximum Lag for ACF Plots*: máximo retardo en las gráficas de la función de autocorrelación.
 - *Maximum Lag for PACF Plots*: máximo retardo en las gráficas de la función de autocorrelación parcial.
 - *Number of Lags for Portmanteau Test*: número de términos que queremos que el sistema use para el test de bondad de ajuste.
 - *Maximum Iterations*: introducir el número máximo de iteraciones en el proceso de estimación.
 - *Stopping Criterion 1*: usamos este campo para introducir el valor que el sistema usa para parar, teniendo en cuenta el cambio en la suma de cuadrados residual.
 - *Stopping Criterion 2*: usamos este campo para introducir el valor que el sistema usa para parar teniendo en cuenta el cambio en los parámetros estimados.

Otras opciones del sistema son:

- *Time sequence plot*: dibuja la serie original o diferenciada versus su índice.
- *Plot ACF*: dibuja la función de autocorrelación de la serie original o diferenciada.

- *Plot PACF*: dibuja la función de autocorrelación parcial de la serie original o diferenciada.
- *Estimate Model*: si la estimación tiene éxito el sistema muestra una tabla con los parámetros del modelo estimados, los errores estándar, los t-valores, los p-valores, la suma de cuadrados residual, un estimador de la varianza del ruido blanco y un test de bondad de ajuste. Si nos escapamos de esta opción tenemos las siguientes opciones adicionales:
 - *Summarize Residuals*: descriptivas de los residuos.
 - *Plot Residuals*: gráfico de normalidad para los residuos.
 - *Plot Residual ACF*: función de autocorrelación para los residuos.
 - *Plot Residual PACF*: función de autocorrelación parcial para los residuos.
 - *Plot Forecasts*: gráfico con las predicciones.
 - *Save Results*: guarda los resultados en un fichero. Tenemos:
 - Forecast Matrix*: matriz de predicciones.
 - Coefficient Summary*: descripción de los coeficientes.
 - Coeff. Corr. Matrix*: matriz de coeficientes de correlación.
 - Residuals*: residuos.

EJERCICIOS

1. Estudiar la serie temporal dada por la variable: Ventas de ordenadores durante 1990-92 en miles de millones.

1413, 1432, 1451, 1497, 1542, 1594, 1613, 1644, 1696, 1714, 1789, 1853, 1862, 1947, 2063, 2192, 2218, 2322, 2323, 2345, 2347, 2413, 2477, 2478, 2512, 2557, 2623, 2672, 2698, 2732, 2793, 2815, 2847, 2893, 2927, 2998.

2. Desde 1988 el Ministerio correspondiente elabora la Encuesta Laboral de Construcción en la que se construyen unos índices de periodicidad trimestral con base en 1988. Entre las obras realizadas por las empresas de construcción, se distinguen obras de edificación y de ingeniería civil. Los índices de ambas series para el periodo 88-93 se encuentran en la tabla adjunta. Analizar ambas series temporales y compararlas.

Año	Trim	Ing. Civil	Edificación
1998	1	77.1	90.9
	2	92.8	98.4
	3	103.2	102.3
	4	126.9	108.5
1989	1	105.1	110.0
	2	125.5	118.4
	3	138.2	117.6
	4	146.4	127.3
1990	1	124.8	111.3
	2	155.2	119.5
	3	160.3	122.6
	4	171.5	137.8
1991	1	144.0	124.8
	2	169.3	132.2
	3	163.2	124.1
	4	170.5	135.1
1992	1	150.1	127.0
	2	156.5	126.5
	3	145.7	121.6
	4	144.5	131.1
1993	1	119.0	114.3
	2	135.1	122.1
	3	143.2	119.6
	4	151.8	136.9